Virginia

Standards of Learning Assessments

Virginia Technical Report

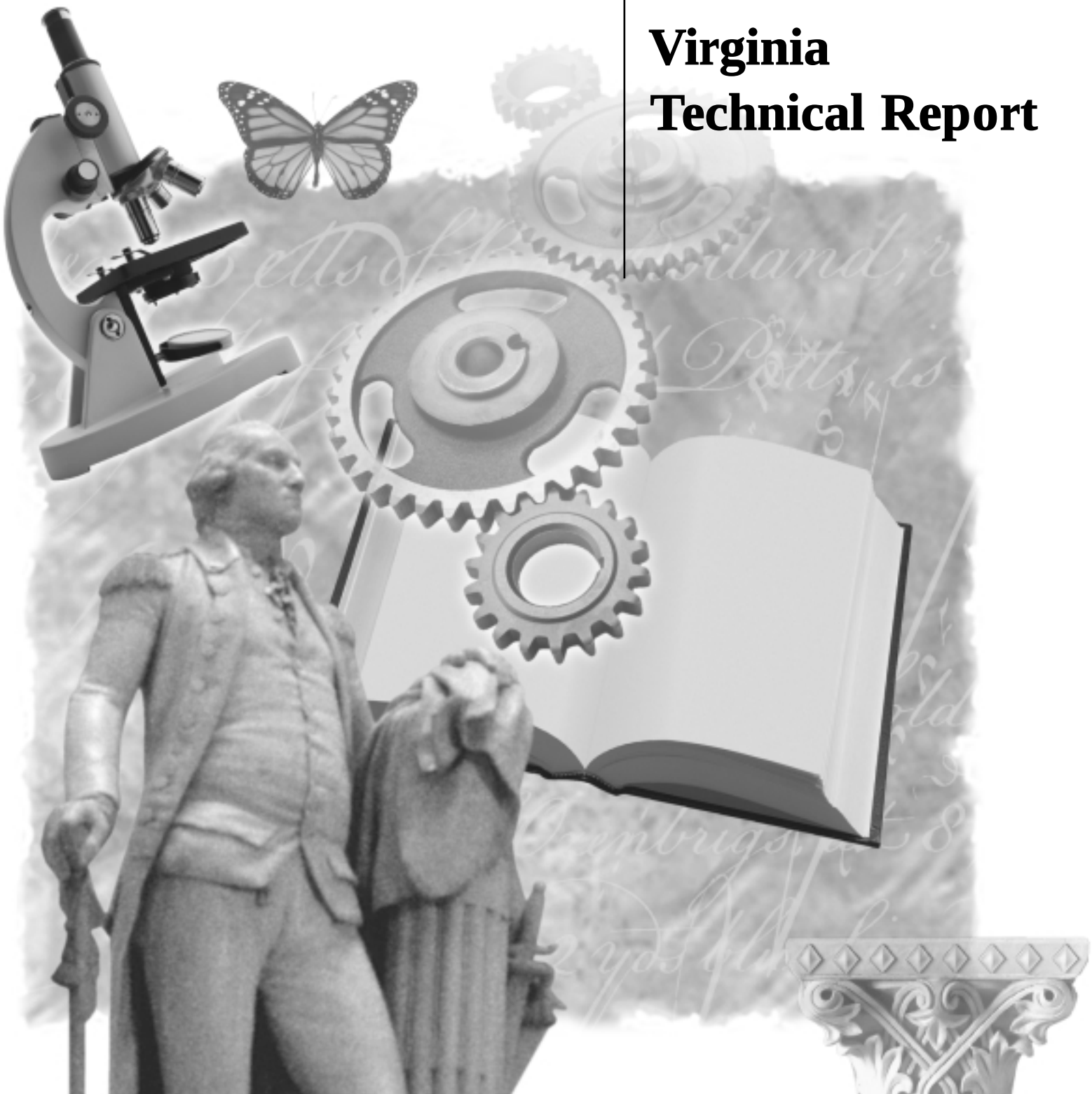

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The purpose of the *Virginia Standards of Learning* (*SOL) Assessment Technical Report*[1] is to inform users and other interested parties about the development and content of the Virginia *SOL* assessments. This Technical Report describes the test development that began in October 1996.

In 1995, the Board of Education of the Commonwealth of Virginia took an important step to raise the expectations for all students in public schools by adopting new *SOLs* in the areas of English, mathematics, history and social science, science, and computer/technology. The Virginia *Standards of Learning* set reasonable targets and expectations for what teachers were expected to teach and what students were expected to learn. These academic standards were used to inform parents and teachers of what students were learning and to make schools accountable for teaching the content found in the *Standards of Learning*. To this end, the Virginia Department of Education (VDOE), in collaboration with hundreds of educators across the Commonwealth and with Harcourt Educational Measurement, developed a series of tests to measure student achievement against the standards.

[1] The appendices for the *Virginia Standards of Learning (SOL) Technical Report* are in two volumes: Appendices A through G are in Book 1, while Appendices H through K are in Book 2.

1. DEVELOPMENT OF THE 1998 *STANDARDS OF LEARNING* ASSESSMENTS

The 1998 *SOL* assessments were composed of multiple-choice items and writing prompts designed to test all the content of all the *SOLs* except where noted on the assessment blueprint (see Section 2.1). Although it was not possible to include items that tested student knowledge on every *SOL* on a single assessment, items were constructed for potential use that did address every *SOL* for subsequent assessment forms. The availability of items provided the potential for assessing an *SOL* in a targeted content area that can be measured using a multiple-choice or writing format[2]. Not all *SOLs* were assessed. See the blueprints for those that are excluded.

1.1 Overview of the *Standards of Learning* Assessments

Students in grades 3, 5, 8, and high school were tested using multiple-choice *SOL* assessments in the content areas listed in Table 1.1. In addition, students in grades 5 and 8, and high school, were assessed using the writing prompt. The *SOL* assessments were cumulative at the elementary and middle-school levels. That is, a content area test at one grade level contained items that addressed *SOL* content from prior grades. For example, grade 5 students taking the Science test encountered items covering content taught in both fourth- and fifth-grade science. Similarly, a grade 8 student taking an *SOL* assessment in Mathematics may have been questioned on mathematics content taught at grades 6, 7, and 8. High school tests were designed to address specific course content, regardless of the grade of the student being tested. More specific information about the *SOLs* covered by each test can be found in the assessment blueprint for the test (see Section 2.1).

Table 1.1 Virginia *Standards of Learning* Assessments at Each Grade Level

Grade 3	Grade 5	Grade 8	High School
1.English: Reading/Writing	1. English: Reading/Literature and Research	1. English: Reading/Literature and Research	1. English: Reading/Literature and Research
2. Mathematics	2. English: Writing	2. English: Writing	2. English: Writing
3. History and Social Science	3. Mathematics	3. Mathematics	3. Algebra I
4. Science	4. History and Social Science	4. History and Social Science	4. Geometry
	5. Science	5. Science	5. Algebra II
	6. Computer/Technology	6. Computer/Technology	6. World History to 1000 A.D./World Geography
			7. World History from 1000 A.D. to the Present/World Geography
			8. United States History
			9. Earth Science
			10. Biology
			11. Chemistry

[2] Not all *SOLs* are assessed. See the assessment blueprint for those that are excluded.

1.2 Responsibility for the Development of the *SOL* Assessments

The creation of the 27 *SOL* assessments needed to assess student learning was a complex and time-consuming undertaking requiring the talents of individuals from the Virginia Department of Education (VDOE), Harcourt Educational Measurement, and local school divisions and local education agencies (LEAs). Teachers, administrators, and content specialists from all over Virginia were recruited to participate in the test development process.

Committee members came to Richmond on several occasions to do the actual work. Follow-up activities were accomplished by Harcourt Educational Measurement in San Antonio, Texas, and by the Virginia Department of Education in Richmond. Table 1.2 shows the groups who assumed the major responsibility in developing the *SOL* assessments.

Table 1.2 Responsibility for the Development of the *SOL* Assessments

Step in Development	Primary Responsibility
• Development of Preliminary Blueprints and Item Specifications	Harcourt; Content Committees
• Development of Preliminary Writing Rubrics	Harcourt; VDOE
• Item Writing	Harcourt
• Item Review	Content Committees
• Construction of Field Test Forms	Harcourt; VDOE
• Pre-Field Test Training Workshops	Harcourt; VDOE; LEAs
• Field Test Administrations	Harcourt; VDOE; LEAs
• Item Data Review	Content Committees
• Bias Review of High School Tests	Bias Review Committees
• Construction of Operational Test Forms	Harcourt; VDOE
• Review of Operational Test Forms	Content Committees; VDOE
• Modification of Special Forms	Harcourt; VDOE
• Review of Special Forms	Special Forms Focus Group (Region 4); Texas Education Service Center
• Final Construction of Operational Forms	Harcourt; VDOE
• Setting Standards for the 1998 *SOL* Assessments	Standard Setting Committees for the Virginia *Standards of Learning*

1.3 Involvement by Virginia Educators

Teachers, administrators, content specialists, and citizens from a variety of locations across Virginia participated in the development of the *SOL* assessments. The efforts of these individuals were crucial in the review of test items and the forms to ensure that the tests adequately measured student knowledge of the content of the *SOL* fairly and without bias.

Assessment Policy Advisory Committee

Members of the Assessment Policy Advisory Committee reviewed and advised the VDOE on the development and implementation of major policies of the *SOL* assessment program. This committee developed recommended guidelines and accommodations for students with disabilities and limited English proficiency. These recommendations were presented to, and adopted by, the Board of Education.

Content Review Committees

The role of the Content Review Committees was to ensure that the assessments matched the *SOL*s, were of appropriate difficulty, and were fair. Committee membership represented all levels of education, from elementary to post secondary, and from all geographic areas of the Commonwealth. Members were Virginia educators who are specialists in the content area for which the items were written or experts in test construction or measurement. The groups were representative of the ethnic and social diversity of Virginia students. The educators' understanding of Virginia curriculum and their extensive classroom experience made them a valuable source of information when developing and reviewing test blueprints, test items, and test forms. The responsibility of these committees was to take a holistic view of the test forms to ensure fairness and a balance of content across reporting categories.

Bias Review Committees and Special Forms Review Focus Group

In addition to the bias review that took place in the Content Review Committees, a separate Bias Review Committee was responsible for examining each item on the high school tests for indications of bias that would impact the performance of an identifiable group of students. Committee members were encouraged to discuss and, if necessary, reject items based on potential gender, ethnic, religious, or geographical bias.

The purpose of the Special Forms Review Focus Group was to examine the forms of the *SOL* assessments that were developed specifically for students with visual disabilities. Committee members were responsible for judging the appropriateness of the test format and editing or deleting items that were inappropriate for students with specific disabilities. In some instances, the only difference was in the size of the print used to accommodate students with visual impairments. In other cases, test forms were constructed for Braille-reading students or for students who required an audio tape of the test to participate in the testing program.

Braille and large-print versions of the test forms were constructed to accommodate students with visual impairments. Audiocassette tapes also were prepared for the Braille and large-print forms, plus the regular test forms.

Report Development Focus Groups

Eight meetings were held across Virginia to collect information from local school personnel on reporting *SOL* assessment results. Representatives from all levels of the LEAs were invited to contribute ideas concerning the type of information and report format that would maximize the usefulness of the information resulting from the test administration.

1.4 Security of Test Materials

Test materials were maintained in locked storage locations when not under supervision of Harcourt Educational Measurement or VDOE personnel. Prior to working with secure test materials, committee members were required to sign Non-Disclosure/Conflict of Interest Agreements. By signing the agreements, participants agreed not to reveal any information about test content, items, scoring keys, or other test-related materials. They also agreed not to reproduce any test materials or use any test-related information for financial gain.

A copy of the non-disclosure agreement is shown in Appendix A.

2. ASSESSMENT DEVELOPMENT AND FIELD TESTING

2.1 Designing Assessment Blueprint and Item Specifications

In order for the new assessments to accurately measure the content of the *Standards of Learning*, Harcourt Educational Measurement staff reviewed the Virginia *SOLs* and developed proposed assessment blueprints for each grade and content area.

Assessment blueprints functioned as maps, or plans, for test constructors. On a blueprint, the identification of content or reporting categories for each *SOL* made it possible for items to be included on a test that matched specific test content. In addition, *SOLs* that could not appropriately be tested by a multiple-choice item format were identified and excluded from testing. Test blueprints also made it possible to determine the relative emphasis given to a content area by calculating the number of test items included in each reporting category. Content Review Committees determined which *SOLs* were to be tested and which could not be tested using multiple-choice format. The test blueprints provided the structure for constructing test forms. Those *SOLs* to be tested were grouped into similar content reporting categories. In many instances, reporting categories were identical to the clustering of standards in the *SOL* documents. At other times, Harcourt Educational Measurement staff members identified reporting categories through a content analysis of the standards.

In December 1996, the Content Review Committees reviewed and modified the draft test blueprints. The committees were organized into grade-specific groups and, at the high school level, into subject-specific groups, to most efficiently judge the grade and content appropriateness of the blueprints. Committee members were afforded the opportunity to revise the number of items in each reporting category in a content area to better reflect the emphasis that they believed a reporting category should have on a particular test. Once approved by committee members, the draft blueprints were used as guides in the development of *SOL* field tests.

Item specifications were general rules or guidelines for the format and layout of test items and ensured a consistency across tests and content areas in the *SOL* assessments. For example, one specification was that all multiple-choice items have four possible choices. Harcourt Educational Measurement assessment development specialists drafted item specifications for each content area and grade level. The specifications provided item writers, item reviewers, and other Harcourt Educational Measurement staff with the guidelines necessary to produce high-quality items tailored to the needs of the *SOL* assessments.

Appendix B of the *Technical Report* contains the advance copies of the assessment blueprints that were published by the Virginia Department of Education. For grades 3, 5, and 8, the assessment blueprints for all content areas within a grade are in the same booklet. For the high school assessments, there are separate blueprint booklets for Secondary English, Algebra I, Geometry, Algebra II, World History to 1000 A.D., World History from 1000 A.D. to the Present, United States History, Biology, Chemistry, and Earth Science. Each booklet introduces the purpose and organization of the *SOL* blueprint, provides development guidelines for the assessment in question, and references the *SOL* assessment content to the *Virginia Standards of Learning* in both tabular and expanded form.

2.2 Developing and Reviewing Test Items

Multiple-Choice Item Development

Upon completion of the item specifications, Harcourt Educational Measurement content specialists and item writers constructed thousands of multiple-choice items to these specifications. Working in collaboration with the VDOE, the Harcourt assessment development team facilitated the review of draft multiple-choice items. The committees were divided into subgroups during the item review process to enable members to focus on items written to their areas of expertise. During the pre-review orientation, committee members were educated in the item review process. They were taught to judge items on the basis of their difficulty, clarity, appropriateness, and relevance to the purpose of the test. Reviewers were also directed to critique each item for its interaction with other items, the appropriateness of accompanying artwork, correctness of keyed responses, and plausibility of the incorrect answer choices (distractors). A copy of the guidelines used by the committees appears in Appendix C.

During the item review process, the Content Review Committees were trained to detect potential item bias in the areas of gender, ethnic, religious, socioeconomic, and regional characteristics. Committee members were encouraged to note their concerns about items they perceived as biased in content or format.

As a result of the review process, some items were eliminated from the prospective field test item bank, and others were marked for revision and inclusion at a later date. Review Committee materials are found in Appendix C.

Writing Prompt Development

Harcourt Educational Measurement staff members drafted over 100 potential writing prompts. By December 1996, 36 writing prompts each for grades 5, 8, and 11 were produced for use in the writing assessment. Prompts were written in the form of a question, an issue, or a hypothetical situation. Prompts were appropriate for the grade level being tested in terms of difficulty, interest, and reading level, as determined by a Content Review Committee.

In January 1997, writing Content Review Committees for grades 5, 8, and high school met to review and revise the prompts. Committee members selected 24 prompts at each grade level for inclusion into the pool of potential prompts for the English writing test. Along with the development of the writing prompts, rubrics were developed to student writing samples in three domains: *Composing, Written Expression,* and *Usage and Mechanics*. These domains were identified by members of the English: Writing Committees. There were nine separate scoring rubrics (one for each domain at each grade level), and they were field tested in the spring 1997 *SOL* writing field test.

2.3 Item and Writing Prompt Field Tests: Spring 1997

Field tests of the *SOL* assessments were conducted in spring 1997. Field testing involved administering items to a sample of students across the Commonwealth. The purpose of a field test is to collect information about test items, not about the students who take the test. More specifically, the following list delineates the purposes of the field test:

1. To provide an array of statistical information, such as the percentage of students answering each item correctly, a difficulty rating for each item, and the ability of each item to discriminate between those students who scored well on the test and those who did not. Field test results also helped to identify items that were potentially biased by ethnicity or gender against students who are members of targeted demographic groups. With this information, committee members were able to identify items for exclusion from the operational forms of the tests.
2. To provide information regarding the test administration procedures, including those for assessing students with disabilities. Examiners were asked to comment on directions for administering the standard test, as well as tests administered with accommodations, such as Braille, large-print, and audio tape forms of the tests.
3. To provide representative teachers, students, and administrators across Virginia with an opportunity to become familiar with the format and general administration procedures of the tests.

The spring *SOL* field tests were administered to provide information about the newly developed test items to the staff at Harcourt Educational Measurement and members of the Content Review Committees. The information provided by the field tests enabled all parties to make informed decisions about test items and the construction of test forms.

Field Test Form Construction

To ensure that sufficient high-quality test items would be available for the two required test forms for the spring 1998 operational assessment, approximately 4,875 items were included in 135 (approximately 5 for each content area) field test forms. Only items that were acceptable to members of the item review committees were included.

Each form was developed to closely reflect the specifications of its test blueprint and consisted of one content area per grade level. Each form within a content area had approximately 30% of its items in common with the other forms. Forms consisted of 28 to 45 unique items and 12 to 18 common or "linking" items. This common-item test design provided the link used to place the difficulty estimates for all the items in each subject area at each grade level on a common scale. The writing assessments were also field tested in spring 1997. Twenty-four different writing prompts for the writing component of the *English: Writing Test* were field tested at grades 5, 8, and 11.

Test Administration Preparation and Materials

Pre-test workshops for representatives of all local school divisions were held across the state prior to the field test. The workshops provided participants an overview of the test content, security expectations, procedures for completing answer documents, and the receipt, distribution, and return of materials.

Three manuals were developed for the *SOL* tests. A *Division Director of Testing Manual*, *School Coordinator's Manual,* and *Examiner's Manual* provided information about the receipt, distribution, security, and return shipment of test materials. In addition to the manuals, directions for administering each *SOL* test were developed and distributed. Several of the *SOL* tests

required the use of ancillary materials such as calculators, protractors, compasses, and rulers. A list of these materials can be found in Table 2.1.

Field Test Administration: Spring 1997

In spring 1997, every student in grades 3, 5, 8, and 11 was involved in field testing the *SOL* assessments in specified content areas. Field test forms were distributed across Virginia to sample a large enough group of students to ensure that the information collected from their responses would allow for analysis of item data. The aim of the sampling procedure was to obtain a representation of students that would mirror the overall composition of Virginia.

A student did not take the full complement of tests, but generally one field test in a content area. For example, students in one third-grade class in a school may have taken a Science field test, while third-grade students in a second class in the building took a Mathematics field test.

In the spring 1997 field test for high school students, some field tests were administered to students who had not taken the course. The scores of the students were eliminated when statistics were run.

Field test administration materials and procedures mirrored those of the operational tests as closely as possible. Separate answer documents incorporating many of the features of the operational answer documents were used to collect demographic data and other information necessary to analyze the results of the field test. Wherever possible, the test forms were modeled on the test blueprints with regard to the number of items and administration time, so that they closely resembled the operational test forms. The major exception occurred with the Reading and Writing tests that relied on passages. Since it was assumed that many items would be rejected after the field test data were analyzed, several more items were included with each reading passage than actually would be used during operational testing.

Twenty-four potential writing prompts were field tested at each of the three grade levels. The number of participants ranged from 266 at grade 11 to 938 at grade 8. The writing samples at each grade level were scored by different teams of readers. Prior to scoring the responses to each prompt, the scoring teams reviewed the rubric and discussed approximately 10 randomly selected writing samples from the field test papers. The scoring process included two blind scorings by team readers with score discrepancies resolved by the team leader.

Field Test Statistics

The descriptive statistics were derived from the spring 1997 field test for each content area, form, and reporting category. They included raw scores, means, and standard deviations by demographic characteristics, form, and reporting categories. The demographic variables included grade level, gender, ethnicity, limited English proficiency status, disability status, and special test accommodations status.

Results from the field test administration that provided a basis for including items in the operational test forms and constructing equivalent forms included item statistics for multiple-choice items and forms, item statistics for the writing prompt domain scores, Rasch item statistics, and differential item functioning (DIF) statistics.

The statistics calculated from the multiple-choice items included:

- numbers of students tested;
- traditional difficulties (*p*-values);
- item-option response distributions for all respondents, for high-, middle-, and low-ability groups, and by gender and ethnic group;
- biserial and point-biserial correlations.

Statistics computed on the results of the writing field test included:

- numbers of students tested;
- frequency distributions, means, and standard deviations for the writing domain raw and total scores;
- correlations between grades and among the multiple-choice and writing domain raw scores;
- percent agreement tables for the writing domain scores assigned by the readers.

The descriptive statistics for the writing domain scores also included analyses by gender and ethnicity. Readers were also asked to perform a qualitative analysis of the writing responses. This analysis is described in more detail below.

To supplement the traditional statistics, item difficulty parameter estimates based on *Item Response Theory* (IRT) were computed. Using this technique, a common underlying construct was assumed to be measurable and estimable as a function of item or test performance, making it possible to estimate item difficulty and item fit.

Differential item functioning (DIF) statistical procedures such as the Mantel-Haenszel Alpha were used to compute the probability that one demographic group is more likely to answer an item correctly than another group. This information was useful in reviewing items and tests for potential bias. High values of the Mantel-Haenszel Alpha indicated that an item interacted differently among equally able students in the reference and comparison groups. When the probability was significantly different across groups, the item warranted further examination. The Mantel-Haenszel Alpha procedure was used to compare white and African-American students, white and Hispanic students, and male and female students. Mantel-Haenszel group differences that exceed a chi-square significance level of 0.10 were “flagged” for further scrutiny.

A Rasch IRT method of computing DIF statistics was also employed to provide item difficulty estimates among demographic groups. Under the assumptions of the Rasch model, the only reason for differences in item difficulty statistics among groups was some group characteristic other than achievement. When the Rasch item difficulty estimates were statistically significant between groups, it was an indicator that further examination was warranted. The Rasch procedure was used to compare white and African-American students, white and Hispanic students, and male and female students. Rasch item difficulty differences exceeding 0.52 were “flagged” for further scrutiny.

A detailed description of methods for identifying DIF in test items can be found in Camilli and Shepard (1994). Wright and Stone (1979, p. 192-195) provide a derivation of the criterion used to flag Rasch item difficulty group differences.

2.4 Writing Prompt Selection and Scoring

Final Selection From Field-Tested Writing Prompts

During the scoring process for field-tested prompts, scorers and team leaders recorded their observations about student responses to each prompt. Subsequently, team leaders were responsible for compiling a qualitative report which addressed the following questions:

- Did the students understand what was being asked of them by the prompt?
- Did the students seem engaged by the prompt?
- Were the students able to effectively focus on a central idea, provide specific information and details, and the like?
- Did the scorers, based upon reading hundreds of student responses to the prompt, recommend that this prompt be used for live testing?

The same prompt was administered to all three grade levels. Papers resulting from this prompt were used by committees to finalize the rubric before the remainder of the prompts were scored. The results of these analyses, in combination with the field-test statistics generated by Harcourt Educational Measurement, were reviewed by the English Writing Committees as they considered which prompts should be included in a prompt item bank for future operational administrations of the *SOL* writing assessment.

Scoring Student Writing Samples: Selecting and Training Scorers

All scoring was done outside the state of Virginia by highly qualified, experienced readers. These readers were drawn from a database of over 1000 college graduates who had completed the selection process for readers. Readers for the Virginia *SOL* writing test had a minimum of a bachelor's degree in an appropriate academic discipline (e.g., English, education), demonstrated ability in performance assessment scoring, and preferably had teaching experience at the elementary or secondary level. The selection process required that each candidate successfully complete a personal interview, a scoring screening sample, a writing sample exercise, and a grammar test. Throughout the selection process, the need for ethnic and racial diversity was emphasized.

The training of readers was conducted by a Performance Assessment Specialist and team leaders, and was critical to high-quality, consistent, and reliable scoring of the *SOL* writing assessments. Readers underwent separate training for each writing prompt. The writing samples used for training scorers were identified from the samples scored during the rangefinding process (see below). These and other writing samples identified by Harcourt Educational Measurement staff and VDOE staff were annotated for use as scoring guides during reader training, qualifying, and calibration. The primary goal of training was to convey to readers the decisions made during rangefinding and to help them internalize the scoring protocol so that they might effectively apply those decisions.

Prospective scorers were provided an opportunity to qualify as a table leader. Table leaders were responsible for supervising small groups of readers and possessed the leadership and communication skills needed to function in a project of this nature. Candidates for table leader positions qualified by achieving a 70% or better exact agreement on each domain when scoring on one set of 10 qualifying papers and 60% or better exact agreement (spring 1998 only) on a second set of papers.

Reader training and qualifying followed the same process as the table leader training and qualifying. The criteria for readers were the same as for table leaders except that some readers who were close to qualifying (e.g., 60% agreement on two sets of papers, spring 1998 only) were permitted to read on probation.

Training began with a discussion of the three writing domains used in the scoring model: composing, written expression, and usage/mechanics. Trainees were introduced to the writing prompt, and then domain-specific training began with a discussion of the features of a domain as well as the score scale. The scale consisted of four score points:

- 4 = Consistent control;
- 3 = Reasonable control;
- 2 = Inconsistent control; and
- 1 = Little or no control.

Following the discussion of each domain and score, prospective table leaders and readers independently scored the domain in a set of papers. Once all domains had been discussed and all domain-specific training sets scored, table leaders and readers began scoring three mixed-domain sets of papers.

To ensure accuracy in scoring, trainees were instructed and practiced scoring regular student responses and a set of calibration prompts each day. Calibration was a process whereby readers re-scored five student papers that previously had been scored by expert scoring team leaders. Calibration sets of student writing samples were dropped in at varying times during the day so scorers were not aware of when they were scoring calibration papers. Scorers who were not consistent with the scores of the experts on the calibration samples were re-trained to improve the accuracy of their scoring. Results of these calibration exercises were reported to the VDOE on a daily basis.

Selecting Anchor Papers

In an exercise described as *rangefinding*, team leaders at Harcourt Educational Measurement familiar with the *SOL* assessment writing prompts organized student writing samples into sets representing high-, middle-, and low-quality responses. The rangefinding process was conducted for each grade level tested. The sets of responses then were used by members of the English Writing Committees to identify model writing samples for each of the three quality levels. These model samples are referred to as *anchor papers* and the identification process as *anchor pulling.*

Anchor pulling involved the scoring of student responses by committee members at each grade level, core members (participants in anchor pulling for all three grade levels), and representatives from Harcourt Educational Measurement and National Computer Systems (NCS), the

subcontractor scoring the writing. During the anchor-pulling process, readers scored the papers independently, the range papers were discussed, and consensus was reached on where the papers fell in the range of scores for a category. Participants checked the range of scores at each quality level to ensure there was no overlap between levels. The anchor-pulling exercise took place over three days, with the focus on one writing domain per day.

Scoring Student Writing Samples

The actual scoring of the student writing responses was carried out by a cadre of trained scorers under the direction of room directors at Harcourt Educational Measurement's Performance Assessment Scoring Center (PASC) in San Antonio. The primary responsibility of the room director during the actual scoring of papers was to ensure high quality scoring and resolve questions that arose during the scoring process. All invalid (unscorable) papers were reviewed by the director to confirm the decision of the scorer. Room directors were also responsible for evaluating readers' performance on the calibration sets. The directors and training supervisor, in conjunction with VDOE staff, monitored reading rates, accuracy rates, and the overall reliability and consistency of scoring. It was also the director's responsibility to re-train readers when necessary.

Prior to the actual scoring, readers were given instruction to cull any papers that were written on the alternate prompt. Scorers also were asked to mark certain papers as "blank" or invalid, including blank papers, off-topic papers, or papers written to the wrong prompt. Readers also were instructed to alert papers that contained troubling content, as well as papers where it appeared that students had cheated or where there had been teacher interference.

2.5 Item Data and Item Bias Reviews: Summer/Fall 1997

Item Data Review

The purpose of the item data review meetings was to conduct a final examination of the items prior to their inclusion in the *SOL* item bank. The item bank, maintained by Harcourt Educational Measurement, served as the repository from which to draw items for current and future forms of the *SOL* assessments. Subsequent to the field test, the Content Review Committees met once again to review items for fairness and bias. The item statistics that were reviewed by the Committees included the Mantel-Haensel Alpha and Rasch item difficulty group differences described above. Committee members were instructed in the interpretation of item statistics and their use in judging the quality and appropriateness of each item in the tests. A sample from the Data and Bias Review Data Books is included in Appendix C.

The data review process provided committee members with an opportunity to discuss concerns about item content, format, bias, and fit with the *SOL*. Participants completed individual rating forms to express their opinion about including an item in the *SOL* item bank. These ratings were tabulated and used to guide decisions about the inclusion of items on the operational test forms. Items that passed all stages of the development process, item review, field test, data review, and bias review were placed in the item bank and were eligible for use on future *SOL* assessments. Item data review materials used by the Content Review Committees are presented in Appendix C.

In addition to reviewing items, draft item specifications and draft blueprints were reviewed by members of the Content Review Committees during the item data review. Committee members offered recommendations for revisions when deemed necessary. Suggested revisions included adjusting the total number of items on the test, adjusting the number and/or type of reporting categories, and adjusting the number of items in each reporting category. The final blueprints were used to construct the first operational test forms, administered in the spring of 1998. Published copies of the blueprints were distributed to all public school teachers in Virginia. Table 2.2 presents, for each of the *SOL* assessments, the numbers of items that were reviewed by the Content Review Committees, and (where available) the numbers and percentages of items that passed the item data review process.

High School Bias Review

Because passing certain high school *SOL* assessments will be a high school graduation requirement, it was especially important that the assessments be free of factors that unfairly impact a group of students. Therefore, a bias review was conducted by a separate Bias Review Committee representing each content area to be tested in addition to the bias review during the data review process. Bias Review Committee members were asked to scrutinize items for potential stereotyping or other forms of bias. The purpose of the bias review was to identify any items that appeared to have the potential to treat any ethnic, gender, or regional group of students differently from other groups. Committee members examined the response distribution for each of the demographic groups identified for the study. The intent of this examination was to determine if members of a certain group were drawn to one or more of the answer choices for the item. If a large percentage of one group selected a particular response, or did not select a particular response, the item was carefully examined.

The training and procedures were similar to those used during the item review meetings. The committee's task focused solely on reviewing test items for potential bias after the items had been reviewed by the Content Review Committees. It was the committee's responsibility to ensure that items were fair to all students and that all students would have an equal opportunity to demonstrate achievement regardless of gender, ethnic background, religion, socio-economic status, or geographic region.

Guidelines used by members of the Bias Committee are presented in Appendix D.

2.6 Review of Operational Forms

Content Review Committees were reconvened in 1998 to review operational forms of the *SOL* assessments. Committee members had the task of approving or editing two forms of each grade level or high school test to determine the content validity and equivalency of the test forms as a whole. While the previous committee reviews were concerned with individual items, the focus of the forms review was the full operational test forms.

Additionally, a Special Forms focus group, in conjunction with staff from the Virginia Department of Education and Harcourt Educational Measurement, met to examine the test items and forms and consider their appropriateness for use on Braille forms, audio tapes, and large-print format.

2.7 Setting Final Standards for the 1998 *SOL* Assessment

As Crocker and Algina (1986, p. 410) point out, "(m)any situations require the setting of cutoff scores before test performance is interpreted. ... The practice of setting cutoff scores is commonly called *standard setting."* In June 1998, the Virginia Board of Education appointed a Standard Setting Advisory Committee (SSAC). The SSAC was responsible for reviewing the procedures and operations of the eight committees involved in the standard setting recommendation process for the 1998 Virginia *Standards of Learning* Tests. Committees were created to set standards for the assessments in grade 3, grade 5, Reading, Writing, Mathematics, History, Science, and Computer/Technology. The assignment of the *SOL* assessments to the eight committees is shown in Table 2.3.

Each of the committees was responsible for setting two cutoff scores for the *SOL* assessments. These cut scores were used to establish three performance categories:

- *Advanced Attainment of the Standards* (Pass)
- *Proficient Attainment of the Standards* (Pass)
- *Does Not Meet the Standards* (Fail)

Two standard setting methods were used to set the cut scores. The method used in the multiple-choice *SOL* assessments is known as the *modified-Angoff* procedure, while that used for the English: Writing assessments at grades 5, 8, and End-of-Course is known as the *Bookmark* procedure. The Bookmark procedure was used for setting standards on the English: Writing assessments, since those assessments made use of both multiple-choice items and a direct writing prompt

The initial steps of the procedures were much the same. In each case, the standard setting committee members were presented with a general definition and description of standard setting as a being a systematic way of making a professional judgment about how many points a student must earn in order to meet a specified criterion.

Next, the committees took the test on which the cut scores were to be set in order to simulate the experience of students taking the test. Only the multiple-choice components of the assessments were taken. For the English: Writing assessments, committee members were not asked to write a paper but were trained briefly in how the writing papers were scored. This training included looking at the scoring guide or rubric, as well as looking at student papers which exemplified each of the score points.

The committee members then were asked to discuss and develop definitions and descriptors of the three performance categories. The purpose of this task was for the committee members to define the particular skills and knowledge that separate those students who are barely proficient in the particular content standards from those who do not meet the content standards. In a similar way, the committee members were asked to define the skills and knowledge separating the students who are advanced from those who are proficient in the content standards.

After these initial steps. the modified-Angoff procedure proceeded as follows:

- Given a copy of the SOL assessment in the content area, committee members were asked to independently examine each of the items. They were asked to estimate the percentage of

barely proficient students who would correctly answer each question correctly. Committee members were instructed to think of what they should be able to do, rather than what they can do now. The procedure was repeated for the advanced category. At the end of this round of ratings, each member had recorded two estimated percents for each question on the assessment.

- Each member's *barely proficient* ratings were averaged and multiplied by the number of the items on the test in order to produce a cut score. The process was repeated for each member's *advanced* ratings.
- The range of the cut scores was presented to the entire committee and discussed. The members had the opportunity to refine their original definitions and descriptors in light of this feedback. When they had completed their discussion, the process started over. All in all, there were three rounds of ratings followed by discussions.
- The end of the final round, the committee's task was completed, and the results of their work was presented to the Board of Education as ranges of potential cut scores.

The Bookmark method differed from the modified-Angoff method in how ratings were obtained from the committee members:

- The committee members were presented with booklets containing the multiple-choice items ordered from easiest to hardest based on the spring 1998 assessment. The booklets were ordered so that the easiest item was at the front of the booklet and the hardest item was at the rear. Interspersed throughout the book were student writing papers ordered from low score point to high score point.
- The members were asked to move through the ordered booklets and to think about the skills and knowledge exemplified by the multiple-choice questions and the scores assigned to the writing prompts. The committee was asked to place a "bookmark" in the booklet at the point where the items and papers prior to the bookmark exemplified the knowledge and skills needed by a student to be considered *barely proficient* in writing. In the same way, a second bookmark was placed by the committee to indicate the knowledge and skills needed by a student to be considered *barely advanced*.
- The committee was provided with a table of each member's ratings and allowed the opportunity to discuss the results, and to refine the definitions and descriptors of the performance categories. When they had completed their discussion, the process was repeated for a total of three rounds of ratings and discussions
- The end of the final round, the committee's task was completed, and the results of their work were presented to the Board of Education as ranges of potential cut scores.

One measure of how well the committees did their work is to examine the convergence of their ratings over the three rounds of the standard setting process (cf. Reckase, 2000, p. 39). That is, as the committee members proceeded with the standard setting process, one would expect that the members would use the feedback given to them to reduce the variation in their ratings. A commonly used index to describe the variation of measurements is the standard deviation, and the expectation would be that, for a given cut score, the standard deviations a committee's ratings would decrease from the initial round of ratings to the final round. Table 2.4 shows that, for the

most part, this was in fact the case. The standard deviation of each committee's ratings decreased from the initial round to the final round of ratings for the proficiency cut score. For the advanced cut score, 23 out of 27 standard settings showed the standard deviations of the committee's ratings decreasing from the initial round to the final round. All of the standard deviations for the ratings at grades 3, 5, and 8 decreased. The standard deviations of the ratings for Algebra I, Earth Science, and Chemistry remained the same, while the ratings for World History from 1000 A.D. to the Present/World Geography increased slightly. Overall, these data suggest that, while the committees were able to use the ratings feedback in setting their standards, they were not dominated by peer pressure to confirm to a single standard.

As was stated above, the results of the committees were presented to the Board of Education. Specifically, the results were presented as a range of suggested cut scores that the Board could take into consideration in setting the final cut scores for the Virginia *SOL* assessments. The Board of Education's final cut scores *SOL* assessments are shown in Table 2.5. The percentages of students failing, and passing at the proficient and advanced levels as a result of applying these cut scores to the spring 1998 *SOL* administration, is shown in Table 2.6.

Appendix E provides additional details of the modified-Angoff and Bookmark standard setting procedures, as well as reports and memoranda from Standard Setting Committees for the Virginia *Standards of Learning*. Included in the appendix is the initial report containing the committee recommendations for each 1998 *SOL* assessment by grade and content area. These recommendations also included the names of the committee members and data from each round of the standard setting. These recommendations were supplemented by a report to the Virginia Board of Education Standard Setting Committee containing the backgrounds and demographics of the committee members, summaries of committee evaluations of the standard setting process, reports from the committee chairs, and the final passing scores established by the Board of Education for the 1998 *SOL* assessments.

Table 2.1 List of Ancillary Materials Used In 1998 Virginia Standards of Learning Assessments

Standards of Learning Assessment	Ancillary Materials
Grade 3	
Mathematics	Ruler, scratch paper
Science	Ruler, scratch paper
Grade 5	
Writing	Dictionary & scratch paper for direct writing component only
Mathematics	Ruler, scratch paper, calculator, protractor
Science	Ruler, scratch paper, calculator
Grade 8	
Writing	Dictionary & scratch paper for direct writing component only
Mathematics	Ruler, scratch paper, calculator, formula sheet
Science	Ruler, scratch paper, calculator
High School End-of-Course	
Writing	Dictionary & scratch paper for direct writing component only
Algebra I	Ruler, scratch paper, calculator, formula sheet
Geometry	Ruler, scratch paper, calculator, formula sheet, compass
Algebra II	Ruler, scratch paper, calculator, formula sheet
Earth Science	Ruler, scratch paper, calculator
Biology	Ruler, scratch paper, calculator
Chemistry	Ruler, scratch paper, calculator, Periodic Table of the Elements

Table 2.2 Numbers and Percents of Items Passing Data Review for the Spring 1998 SOL Assessments

Standards of Learning Assessment	No. of Items Reviewed	No. of Items Passing Data Review	% of Items Passing Data Review
Grade 3			
English: Reading	150	140	93
English: Writing[1]	100	-	-
Mathematics	250	230	92
History	320	302	94
Science	200	175	88
Grade 5			
English: Reading/Lit. & Resrch.	250	226	90
English: Writing[1]	200	-	-
Mathematics	250	238	95
History[1]	200	-	-
Science	250	220	88
Computer/Technology	150	146	97
Grade 8			
English: Reading/Lit. & Resrch.	250	241	96
English: Writing[1]	320	-	-
Mathematics	300	275	92
History	250	210	84
Science	200	161	81
Computer/Technology	200	151	76
High School			
English: Reading/Lit. & Resrch.	270	235	87
English: Writing	270	230	85
Algebra I	450	407	90
Geometry	225	172	76
Algebra II	225	209	93
United States History	300	269	89
Wrld. Hist. to 1000 A.D./W. Geog. [1]	300	-	-
Wrld. Hist. from 1000 A.D./W. Geog.	300	278	93
Earth Science[1]	250	-	-
Biology	250	224	90
Chemistry	250	217	87

[1] Number and percents of items passing Data Review unavailable

Table 2.3 Assignment of Standards of Learning Assessments to Standard Setting Committees

	Standard Setting Committee Assignments							
Standards of Learning Assessment	1	2	3	4	5	6	7	8
Grade 3								
English: Reading/Writing	•							
Mathematics	•							
History	•							
Science	•							
Grade 5								
English: Reading/Lit. & Resrch.		•						
English: Writing			•					
Mathematics		•						
History		•						
Science		•						
Computer/Technology				•				
Grade 8								
English: Reading/Lit. & Resrch.					•			
English: Writing			•					
Mathematics						•		
History							•	
Science								•
Computer/Technology				•				
High School End-of-Course								
English: Reading/Lit. & Resrch.					•			
English: Writing			•					
Algebra I						•		
Geometry						•		
Algebra II						•		
United States History							•	
Wrld. Hist. to 1000 A.D./W. Geog.							•	
Wrld. Hist. from 1000 A.D./W. Geog.							•	
Earth Science								•
Biology								•
Chemistry								•

Table 2.4 Initial and Final Standard Deviations of Standard Setting Committee Members' Ratings

Standards of Learning Assessment	No. of Committee Members	Proficient Cut Score Ratings		Advanced Cut Score Ratings	
		Initial SD	Final SD	Initial SD	Final SD
Grade 3					
English: Reading/Writing	19	6.0	4.6	5.0	2.1
Mathematics	19	5.9	4.9	4.3	3.3
History	19	4.5	3.8	5.0	3.3
Science	19	5.5	4.0	4.0	1.9
Grade 5					
English: Reading/Lit. & Resrch.	20	5.1	3.6	3.7	1.8
English: Writing	19	4.6	3.2	2.4	1.7
Mathematics	20	5.6	4.6	3.0	2.2
History	20	4.5	3.8	3.1	2.2
Science	20	4.5	3.6	3.6	1.8
Computer/Technology	11	4.7	1.7	2.0	1.8
Grade 8					
English: Reading/Lit. & Resrch.	17	4.0	3.6	3.0	2.3
English: Writing	19	4.0	2.4	11.1	2.0
Mathematics	19	4.7	3.0	2.5	2.2
History	21	6.3	4.8	4.0	2.8
Science	20	3.3	3.0	3.2	1.8
Computer/Technology	11	6.0	3.2	3.0	2.3
High School End-of-Course					
English: Reading/Lit. & Resrch.	17	3.6	3.4	4.1	3.4
English: Writing	19	7.7	4.3	3.3	2.1
Algebra I	19	3.8	3.4	1.9	1.9
Geometry	19	6.0	2.8	2.1	1.5
Algebra II	19	4.8	3.5	2.3	1.6
United States History	21	7.4	5.7	5.3	3.8
Wrld. Hist. to 1000 A.D./W. Geog.	19	4.3	3.9	3.0	3.5
Wrld. Hist. from 1000 A.D./W. Geog.	20	5.3	4.9	3.9	2.9
Earth Science	20	2.5	2.4	1.4	1.4
Biology	20	3.3	2.6	2.7	2.1
Chemistry	20	2.9	2.3	1.4	1.4

Table 2.5 Virginia Standards of Learning Assessments: Passing Scores Established by the Board of Education

Standards of Learning Assessment	Max. Score	Pass (proficient) Raw Score	Pass (proficient) Percent of Max. Score	Pass (advanced) Raw Score	Pass (advanced) Percent of Max. Score
Grade 3					
English: Reading/Writing	45	32	71%	42	93%
Mathematics	50	36	72	45	90
History	40	24	60	36	90
Science	40	27	68	36	90
Grade 5					
English: Reading/Lit. & Resrch.	42	28	67%	39	93%
English: Writing	44	32	73	41	93
Mathematics	50	34	68	46	92
History	40	26	65	37	93
Science	40	26	65	37	93
Computer/Technology	30	17	57	27	90
Grade 8					
English: Reading/Lit. & Resrch.	42	27	64%	37	88%
English: Writing	44	30	68	41	93
Mathematics	60	37	62	55	92
History	50	33	66	45	90
Science	50	29	58	45	90
Computer/Technology	40	26	65	36	90
High School End-of-Course					
English: Reading/Lit. & Resrch.	42	24	57%	37	88%
English: Writing	54	37	69	49	93
Algebra I	50	27	54	45	90
Geometry	45	27	60	41	91
Algebra II	50	31	62	45	90
United States History	61	40	66	55	90
Wrld. Hist. to 1000 A.D./W. Geog.	61	33	61	55	90
Wrld. Hist. from 1000 A.D./W. Geog.	63	36	57	57	90
Earth Science	50	30	60	45	90
Biology	50	26	52	45	90
Chemistry	50	27	54	45	90

Table 2.6 *SOL* Assessments: Spring 1998 Administration Results

Standards of Learning Assessment	% Fail	% Pass	
		Proficient	Advanced
Grade 3			
English: Reading/Writing	45	44	11
Mathematics	37	39	24
History	51	46	3
Science	37	53	10
Grade 5			
English: Reading/Lit. & Resrch.	32	57	11
English: Writing	35	53	12
Mathematics	53	41	5
History	67	32	1
Science	41	56	3
Computer/Technology	28	62	10
Grade 8			
English: Reading/Lit. & Resrch.	35	50	14
English: Writing	29	59	11
Mathematics	48	45	7
History	65	33	3
Science	29	62	9
Computer/Technology	37	54	9
High School End-of-Course			
English: Reading/Lit. & Resrch.	28	55	1
English: Writing	29	59	11
Algebra I	60	37	3
Geometry	48	48	4
Algebra II	69	28	3
United States History	70	27	3
Wrld. Hist. to 1000 A.D./W. Geog.	38	57	5
Wrld. Hist. from 1000 A.D./W. Geog.	59	38	5
Earth Science	42	53	4
Biology	28	66	6
Chemistry	46	52	2

3. SPRING 1998 ADMINISTRATION: RELIABILITY, VALIDITY, AND DESCRIPTIVE STATISTICS

This section presents a summary of the descriptive statistics and reliabilities for the spring 1998 administration of the *SOL* assessments. This section, together with the *Technical Report* appendices, provides details of the psychometric and statistical analyses performed after the first operational administration of the *SOL* assessments.

In general, analyses are provided for both the writing assessments in grades 5, 8, and end-of-course, and the multiple-choice assessments at grades 3, 5, 8, and end-of-course. For the writing assessments, analyses are provided for each combination of multiple-choice section and writing prompt. Analyses for the multiple-choice assessments are presented for both the Core 1 ("Main") and Core 2 ("Makeup") forms of the assessments.

3.1 Summary of Reliabilities and Scale Score Descriptive Statistics

Tables 3.1 through 3.4 present the raw score statistics and reliabilities for each grade and form of the multiple-choice *SOL* assessments, and include:

- the number of items;
- the numbers of students[3];
- the means and standard deviations of the students' scale scores[4];
- the *Kuder-Richardson Formula 20* (KR20) internal consistency reliability estimate (Crocker & Algina, 1987, p. 139);
- the standard error of measurement;
- the mean raw score as a proportion of the maximum obtainable score; and
- the conditional standard errors of measurement for the proficient and advanced cut scores.

Tables 3.5 through 3.7 present the statistics for the grades 5, 8, and end-of-course writing assessments, and include:

- the specific combination of writing prompt and multiple-choice section that was administered;
- the number of items that were on the writing assessment;
- the maximum obtainable raw score possible for the writing assessment;
- the numbers of students;
- the means and standard deviations of the students' scale scores;

[3] Note the numbers of students reported in these tables may be lower than the totals reported in the statewide summaries. These differences are to inclusion of all student results in the state summaries and the exclusion of incomplete student results in the statistical summaries.

[4] The derivation of the scale scores reported in this section is described in Section 4 and in the Technical Note at the end of this report.

- the *coefficient alpha* internal consistency reliability estimate (Crocker & Algina, 1987, p. 138);
- the standard error of measurement;
- the mean raw score as a proportion of the maximum obtainable score; and
- the conditional standard errors of measurement for the proficient and advanced cut scores.

Tables 3.8 through 3.10 present the correlation mat of the raw scores for each set of multiple-choice SOL assessments in grades 3, 5, and 8. In each table, the intercorrelations for the Core 1 forms of the set of assessments are above the main diagonal, while the intercorrelations for the Core 2 forms are below the main diagonal.

Additional statistical information regarding the multiple-choice and writing assessments can be found in following appendices:

Appendix F provides additional descriptive statistics and frequency distributions of the raw scores and scale scores of the *SOL* assessments. The analyses for the grades 5, 8, and end-of-course writing assessments are presented first, and are followed by the analyses for the grades 3, 5, 8, and end-of-course multiple-choice assessments. These analyses were used to produce the tables in Section 3 of this report.

Appendix G presents the average *p*-values and adjusted Rasch item difficulties. The results are presented by grade for each content area. For each grade, the content area results for the Core 1 form are followed by the results for the Core 2 form.

Multiple-choice item statistics are shown in Appendix H. For each multiple-choice item, the statistics include the *p*-value, point-biserial correlation, Rasch difficulty, standard error, and mean square fit. Within each grade, the results for the assessment content areas are reported in pairs, with results for the Core 1 form followed by the results for the Core 2 form.

Appendix I contains the statistics for the writing prompts and assessments. The analyses of the writing prompts can be found in the *BIGSTEPS* (Linacre & Wright, 1991) Rasch analysis program output files in this appendix. Detailed information is presented regarding the item measures, infits, and outfits. Of special interest is Table 3 of each output, which summarizes person, item, and step measure results.

3.2 The Reliability of Passing Cut Scores: Decision Consistency and Accuracy

Tables 3.11 through 3.17 present the results of a set of analyses that were performed to estimate the accuracy and consistency of decisions based on the cut scores for passing (proficient) on the Virginia *SOL* assessments. These analyses make use of the methods outlined and implemented in Livingston and Lewis (1995), Haertel (1996), and Young and Yoon (1998).

The *accuracy* of a decision is the extent to which it would agree with the decisions that would be made if each student could somehow be tested with all possible parallel forms of the assessment that were used. The *consistency* of a decision is the extent to which it would agree with the decisions that would be made if the students had taken a different form of the examination, equal in difficulty and covering the same content as the form they actually took. Students can be misclassified in one of two ways. Students who were truly below a proficiency cutpoint, but were

classified on the basis of the assessment as being above a cutpoint, are considered to be *false positives*. In a similar fashion, students who were truly above a proficiency cutpoint, but were classified as being below a cutpoint, are considered to be *false negatives*.

For each *SOL* multiple-choice and writing assessment, these tables include:

- the proportion of consistent classifications;
- the proportion of accurate classifications;
- the proportion of false positives;
- the proportion of false negatives.

Note that these tables follow the general rule that decision consistency will be less than decision accuracy.

3.3 Inter-Rater Reliability

Tables 3.18 and 3.19 provide evidence for the inter-rater reliability of the writing assessments. Each writing prompt was read and scored by two independent raters. When the two raters assigned the same score to a student's paper, the scores were said to be in *exact agreement*. Scores that differed by exactly one score point were said to be *adjacent*, while scores that differed by two or more score points were said to be *non-adjacent*. All papers that were non-adjacent were reviewed by the room directors before a final score was assigned.

Each of these tables includes:

- the writing prompt and writing domain score;
- the numbers of students for which the writing domain inter-rater reliabilities were calculated; and
- the percentages of papers that were in exact agreement, adjacent, or non-adjacent.

3.4 Validity

Tables 3.20 and 3.21 provide validity evidence related to the external structure of the assessment by examining the relationship of the *SOL* assessments with the *Stanford Achievement Test, Ninth Edition*, and the *Literacy Passport Test (LPT)*. Specifically, these data address the question "Do schools that score well on the *Stanford 9* or the *LPT* also score well on the *SOL* tests in content areas where there are similar knowledge and skills?" (p. 8, Virginia Department of Education, 1999).

The building-level results in Tables 3.30 and 3.21 show the correlations of school pass rates on the *SOL* tests in English and mathematics with national percentile ranks on the *Stanford 9* and/or pass rates on the *LPT*. The student-level results in Table 3.22 present the correlations of *SOL* raw and scale scores with the respective *Stanford 9* total and subtest raw and scale scores.

As the Virginia Department of Education's (1998, p.8) interpretative report regarding the building-level results states:

> In content areas and grade levels where there were reasonable matches of content … [t]hese data show a strong relationship between the relative standing of Virginia's schools on the SOL tests and both the *Stanford 9* and the *LPT*. While overall performance on the SOL tests is dramatically lower than on the *Stanford 9* and the *LPT*, the relative standing among schools is very similar. Though varying among grades and content areas, schools that scored well on the *Stanford 9* or *LPT* generally scored well on related SOL tests, and vice versa.

Similar results were found for the student-level results. That is, students who scored well on the *Stanford 9* tended to score well on the *SOL* assessment.

The results which are summarized in Tables 3.20 and 3.21 were taken from the interpretive report, which can be found in Appendix J; the results which are summarized in Table 3.22 are taken from Appendix K.

Table 3.1 Virginia SOL Grade 3 Assessments: Scale Score Statistics, Reliabilities, and SEMs

Standards of Learning Assessment	Form	No. of Items	N	Mean	SD	KR20	SEM	Prop. Max.	Conditional SEM	
									Prof. Cut	Adv. Cut
English: Reading/Writing	Core 1	45	80,262	406.6	67.4	0.90	21.3	0.69	18.8	33.2
	Core 2	45	3,934	404.0	65.5	0.91	19.7	0.68	18.8	33.2
Mathematics	Core 1	50	80,262	427.4	88.4	0.91	26.5	0.74	24.6	36.0
	Core 2	50	3,934	429.8	88.1	0.91	26.4	0.74	24.7	36.1
History	Core 1	40	80,262	397.5	48.9	0.84	19.6	0.58	17.8	28.0
	Core 2	40	3,934	396.9	45.3	0.82	19.2	0.58	18.2	28.1
Science	Core 1	40	80,262	415.0	67.7	0.85	26.2	0.70	22.1	32.8
	Core 2	40	3,934	414.0	59.3	0.84	23.7	0.70	22.5	33.0

.

Table 3.2 *SOL* Grade 5 Assessments: Scale Score Statistics, Reliabilities, and SEMs

Standards of Learning Assessment	Form	No. of Items	N	Mean	SD	KR20	SEM	Prop. Max.	Conditional SEM	
									Prof. Cut	Adv. Cut
English: Reading/Lit. & Resrch.	Core 1	42	75,764	424.1	58.7	0.89	19.5	0.72	17.3	30.5
	Core 2	42	3,864	425.4	62.5	0.90	19.8	0.72	18.2	31.0
Mathematics	Core 1	50	75,764	397.4	56.1	0.88	19.4	0.64	17.2	28.5
	Core 2	50	3,864	396.5	57.4	0.89	19.0	0.62	17.3	29.0
History	Core 1	40	75,764	379.6	40.9	0.80	18.3	0.55	17.3	30.6
	Core 2	40	3,864	382.6	43.2	0.82	18.3	0.56	17.4	30.6
Science	Core 1	40	75,764	408.1	44.8	0.81	19.5	0.66	17.2	29.9
	Core 2	40	3,864	413.1	47.2	0.84	18.9	0.66	17.0	29.9
Computer/Technology	Core 1	30	75,764	427.4	50.6	0.81	22.1	0.66	18.5	29.2
	Core 2	30	3,864	429.1	50.7	0.82	21.5	0.68	18.2	34.6

Table 3.3 *SOL* Grade 8 Assessments: Scale Score Statistics, Reliabilities, and SEMs

Standards of Learning Assessment	Form	No. of Items	N	Mean	SD	KR20	SEM	Prop. Max.	Conditional SEM	
									Prof. Cut	Adv. Cut
English: Reading/Lit. & Resrch.	Core 1	42	70,076	423.1	67.7	0.87	24.4	0.69	22.1	31.7
	Core 2	42	3,093	415.0	67.8	0.88	23.5	0.65	21.6	31.2
Mathematics	Core 1	60	70,076	408.7	55.5	0.92	15.7	0.62	13.5	23.1
	Core 2	60	3,093	394.0	51.6	0.92	14.6	0.57	13.6	23.1
History	Core 1	50	70,076	377.7	57.3	0.88	19.8	0.57	18.8	28.5
	Core 2	50	3,093	368.0	49.4	0.86	18.5	0.54	18.6	31.2
Science	Core 1	50	70,076	429.6	49.8	0.88	17.3	0.68	14.9	23.6
	Core 2	50	3,093	416.3	45.1	0.87	16.3	0.63	14.9	23.7
Computer/Technology	Core 1	40	70,076	417.9	59.6	0.86	22.3	0.68	20.1	30.2
	Core 2	40	3,093	400.7	54.2	0.86	20.3	0.62	19.9	30.2

Table 3.4 *SOL* End-of-Course Assessments: Scale Score Statistics, Reliabilities, and SEMs

Standards of Learning Assessment	Form	No. of Items	N	Mean	SD	KR20	SEM	Prop. Max.	Conditional SEM	
									Prof. Cut	Adv. Cut
English: Reading/Lit. & Resrch.	Core 1	42	55,222	434.4	62.7	0.89	20.8	0.67	17.7	26.1
	Core 2	42	2,958	443.4	61.1	0.89	20.3	0.67	17.6	26.1
Algebra I	Core 1	50	68,949	395.7	43.7	0.88	15.1	0.51	13.3	21.7
	Core 2	50	3,830	384.8	34.0	0.82	14.4	0.47	13.3	23.5
Geometry	Core 1	45	49,539	403.7	47.5	0.85	18.4	0.60	15.7	25.8
	Core 2	45	2,572	410.8	51.2	0.88	17.7	0.62	15.9	25.9
Algebra II	Core 1	50	41,056	379.2	50.5	0.86	18.9	0.53	16.8	26.5
	Core 2	50	1,951	371.8	48.3	0.86	18.1	0.51	17.0	29.2
United States History	Core 1	61	55,220	371.3	58.2	0.90	18.4	0.54	16.9	26.4
	Core 2	61	4,734	370.6	57.6	0.91	17.3	0.53	16.8	26.3
Wrld. Hist. to 1000 A.D./W. Geog.	Core 1	61	32,779	415.5	46.6	0.91	14.0	0.60	12.2	19.7
	Core 2	61	1,872	421.1	43.6	0.91	13.1	0.61	12.2	19.7
Wrld. Hist. from 1000 A.D./W. Geog.	Core 1	63	26,212	392.8	47.3	0.91	14.2	0.53	12.8	21.2
	Core 2	63	845	389.3	49.1	0.87	17.7	0.51	12.8	21.1
Earth Science	Core 1	50	54,052	409.2	48.5	0.87	17.5	0.62	15.8	24.7
	Core 2	50	3,651	411.9	49.8	0.87	18.0	0.64	16.0	27.5
Biology	Core 1	50	65,526	425.5	43.4	0.88	15.0	0.62	13.1	20.8
	Core 2	50	4,065	419.9	43.0	0.88	14.9	0.62	12.9	22.9
Chemistry	Core 1	50	40,661	404.5	42.8	0.88	14.8	0.55	13.5	22.0
	Core 2	50	2,785	411.3	42.1	0.88	14.6	0.60	13.5	23.6

Table 3.5 *SOL* Grade 5 Writing Assessments: Scale Score Statistics, Reliabilities, and SEMs

Writing Assessment Configuration										Conditional SEM	
Prompt	MC	No. of Items	Max. Score	N	Mean	SD	Alpha	SEM	Prop. Max.	Prof. Cut	Adv. Cut
Core 1	Core 1	21	44	64,880	422.3	63.2	0.82	26.8	0.76	18.8	29.1
Core 1	Core 2	21	44	3,494	404.5	61.9	0.83	25.5	0.70	18.8	29.6
Core 2	Core 1	21	44	5,717	423.9	61.1	0.81	26.6	0.76	18.8	28.6
Core 2	Core 2	21	44	442	399.0	63.3	0.84	25.3	0.68	19.2	31.5

Table 3.6 *SOL* Grade 8 Writing Assessments: Scale Score Statistics, Reliabilities, and SEMs

Writing Assessment Configuration										Conditional SEM	
Prompt	MC	No. of Items	Max. Score	N	Mean	SD	Alpha	SEM	Prop. Max.	Prof. Cut	Adv. Cut
Core 1	Core 1	21	44	68,153	417.7	47.0	0.81	20.5	0.72	16.0	23.7
Core 1	Core 2	21	44	4,945	407.0	48.5	0.83	20.0	0.69	16.0	24.1
Core 2	Core 1	21	44	4,881	424.4	47.6	0.80	21.3	0.74	16.3	24.1
Core 2	Core 2	21	44	650	417.0	49.5	0.82	21.0	0.72	16.0	24.5

Table 3.7 *SOL* End-of-Course Writing Assessments: Scale Score Statistics, Reliabilities, and SEMs

Writing Assessment Configuration										Conditional SEM	
Prompt	MC	No. of Items	Max. Score	N	Mean	SD	Alpha	SEM	Prop. Max.	Prof. Cut	Adv. Cut
Core 1	Core 1	31	54	50,759	429.0	57.1	0.87	20.6	0.74	16.1	25.4
Core 1	Core 2	31	54	4,841	411.7	55.6	0.88	19.3	0.71	16.1	27.2
Core 2	Core 1	31	54	3,142	426.0	55.0	0.86	20.6	0.75	16.1	27.2
Core 2	Core 2	31	54	216	403.8	49.1	0.84	19.6	0.69	16.1	27.7

Table 3.8 Correlations Among Grade 3 *SOL* Assessments

Standards of Learning Assessment	1	2	3	4
1. English: Reading/Writing	•	.72	.78	.78
2. Mathematics	.78	•	.75	.78
3. History	.77	.73	•	.78
4. Science	.75	.76	.76	•

Note: Core 1 correlations are above the main diagonal; Core 2 correlations are below the main diagonal

Table 3.9 Correlations Among Grade 5 *SOL* Assessments

Standards of Learning Assessment	1	2	3	4	5
1. English: Reading/Writing	•	.72	.71	.76	.72
2. Mathematics	.73	•	.69	.74	.69
3. History	.72	.73	•	.74	.71
4. Science	.78	.76	.75	•	.73
5. Computer/Technology	.75	.70	70	.75	•

Note: Core 1 correlations are above the main diagonal; Core 2 correlations are below the main diagonal

Table 3.10 Correlations Among Grade 8 *SOL* Assessments

Standards of Learning Assessment	1	2	3	4	5
1. English: Reading/Writing	•	.72	.74	.75	.72
2. Mathematics	.72	•	.75	.78	.73
3. History	.73	.72	•	.78	.73
4. Science	.73	.77	.76	•	.74
5. Computer/Technology	.70	.73	.70	.74	•

Note: Core 1 correlations are above the main diagonal; Core 2 correlations are below the main diagonal

Table 3.11 *SOL* Grade 3 Assessments: Decision Accuracy and Consistency Rates[5]

Standards of Learning Assessment	Form	Consistency	Accuracy	False Positives	False Negatives
English: Reading/Writing	Core 1	.87	.91	.05	.05
	Core 2	.87	.91	.04	.05
Mathematics	Core 1	.88	.91	.04	.05
	Core 2	.88	.91	.05	.04
History	Core 1	.83	.88	.06	.06
	Core 2	.82	.87	.07	.06
Science	Core 1	.83	.88	.06	.06
	Core 2	.82	.87	.05	.07

Table 3.12 *SOL* Grade 5 Assessments: Decision Accuracy and Consistency Rates

Standards of Learning Assessment	Form	Consistency	Accuracy	False Positives	False Negatives
English: Reading/Lit. & Resrch.	Core 1	.86	.90	.04	.06
	Core 2	.87	.90	.04	.06
Mathematics	Core 1	.86	.90	.06	.04
	Core 2	.86	.90	.06	.04
History	Core 1	.83	.88	.08	.05
	Core 2	.84	.89	.06	.05
Science	Core 1	.80	.86	.07	.07
	Core 2	.82	.87	.07	.06
Computer/Technology	Core 1	.80	.86	.06	.08
	Core 2	.82	.87	.07	.06

Table 3.13 *SOL* Grade 8 Assessments: Decision Accuracy and Consistency Rates

Standards of Learning Assessment	Form	Consistency	Accuracy	False Positives	False Negatives
English: Reading/Lit. & Resrch.	Core 1	.85	.89	.05	.06
	Core 2	.85	.89	.05	.05
Mathematics	Core 1	.88	.91	.04	.04
	Core 2	.89	.92	.04	.04
History	Core 1	.87	.91	.05	.04
	Core 2	.87	.91	.06	.04
Science	Core 1	.85	.90	.05	.06
	Core 2	.84	.88	.05	.06
Computer/Technology	Core 1	.84	.89	.06	.05
	Core 2	.84	.88	.07	.05

[5] The decision accuracy and consistency estimates in Tables 3.8 through 3.14 were obtained using the methods outlined in Livingston and Lewis (1995), Haertel (1996), and Young and Yoon (1998).

Table 3.14 *SOL* End-of-Course Assessments: Decision Accuracy and Consistency Rates

Standards of Learning Assessment	Form	Consistency	Accuracy	False Positives	False Negatives
English: Reading/Lit. & Resrch.	Core 1	.87	.90	.05	.05
	Core 2	.87	.90	.04	.05
Algebra I	Core 1	.86	.90	.06	.04
	Core 2	.86	.90	.06	.03
Geometry	Core 1	.83	.88	.07	.05
	Core 2	.85	.89	.06	.05
Algebra II	Core 1	.88	.92	.05	.03
	Core 2	.88	.92	.05	.03
United States History	Core 1	.90	.93	.04	.03
	Core 2	.90	.93	.04	.03
Wrld. Hist. to 1000 A.D./W. Geog.	Core 1	.86	.90	.06	.05
	Core 2	.87	.90	.04	.05
Wrld. Hist. from 1000 A.D. /W. Geog.	Core 1	.89	.92	.04	.04
	Core 2	.89	.92	.04	.04
Earth Science	Core 1	.84	.89	.06	.05
	Core 2	.84	.89	.05	.06
Biology	Core 1	.85	.89	.05	.06
	Core 2	.86	.90	.04	.06
Chemistry	Core 1	.85	.89	.06	.05
	Core 2	.85	.89	.06	.05

Table 3.15 *SOL* Grade 5 Writing Assessments: Decision Accuracy and Consistency Rates

Writing Assessment Configuration					
Prompt	MC	Consistency	Accuracy	False Positives	False Negatives
Core 1	Core 1	.82	.87	.07	.06
Core 1	Core 2	.84	.89	.06	.05
Core 2	Core 1	.81	.87	.07	.07
Core 2	Core 2	.85	.89	.06	.05

Table 3.16 *SOL* Grade 8 Writing Assessments: Decision Accuracy and Consistency Rates

Writing Assessment Configuration					
Prompt	**MC**	**Consistency**	**Accuracy**	**False Positives**	**False Negatives**
Core 1	Core 1	.82	.87	.06	.06
Core 1	Core 2	.83	.88	.07	.06
Core 2	Core 1	.81	.86	.08	.06
Core 2	Core 2	.83	.88	.05	.07

Table 3.17 *SOL* End-of-Course Writing Assessments: Decision Accuracy and Consistency Rates

Writing Assessment Configuration					
Prompt	**MC**	**Consistency**	**Accuracy**	**False Positives**	**False Negatives**
Core 1	Core 1	.84	.89	.05	.06
Core 1	Core 2	.85	.89	.06	.05
Core 2	Core 1	.84	.89	.05	.06
Core 2	Core 2	.84	.89	.06	.06

Table 3.18 *SOL* Grade 8 Writing Assessment: Inter-Rater Reliability

		Percent		
Prompt/ Writing Domain Score	**N**	**Perfect Agreement**	**Adjacent**	**Non-Adjacent**
Core 1				
Composing	152,431	75.1	24.8	0.1
Written Expression	152,431	74.2	25.6	0.2
Usage and Mechanics	152,429	69.2	30.6	0.2
Core 2				
Composing	15,460	72.0	27.9	0.2
Written Expression	15,460	72.3	27.6	0.2
Usage and Mechanics	15,460	68.9	30.8	0.2

Table 3.19 *SOL* End-of-Course Writing Assessment: Inter-Rater Reliability

Prompt/ Writing Domain Score	N	Percent		
		Perfect Agreement	Adjacent	Non-Adjacent
Core 1				
Composing	120,925	66.2	33.5	0.3
Written Expression	120,925	64.0	35.5	0.5
Usage and Mechanics	120,925	61.0	38.4	0.6
Core 2				
Composing	5,742	67.5	32.4	0.1
Written Expression	5,742	65.7	34.0	0.3
Usage and Mechanics	5,742	60.7	39.0	0.3

Table 3.20 Building Pass Rates on *SOL* Assessments Correlated with National Percentile Ranks on *Stanford 9* Assessment

SOL Assessment (Spring 1998)	*Stanford 9* (Spring 1997)	Number of Schools	Spearman Rank Order Correlation
Grade 3			
English: Reading & Writing	Total Reading	1,071	.78
Mathematics	Total Mathematics	1,071	.75
Grade 5			
English: Reading/Lit. & Rsrch	Total Reading	1,039	.78
Mathematics	Total Mathematics	1,039	.75
Grade 8			
English: Reading/Lit. & Rsrch	Total Reading	368	.81
Mathematics	Total Mathematics	368	.83
End-of-Course			
English: Reading/Lit. & Rsrch	Total Reading	315	.62
Algebra I	Total Mathematics	312	.53
Geometry	Total Mathematics	308	.71
Algebra II	Total Mathematics	307	.66

Table 3.21 Building Pass Rates on *SOL* Assessments Correlated with National Percentile Ranks on Grade 6 *Literacy Passport Tests*

***SOL* Assessment (Spring 1998)**	***LPT* Grade 6 (Spring 1998)**	**Number of Schools**	**Spearman Rank Order Correlation**
Grade 5			
English: Reading/Lit. & Rsrch	Reading	272	.64
Writing	Reading	270	.68
Grade 8			
English: Reading/Lit. & Rsrch	Reading	288	.75
Writing	Reading	287	.61

Table 3.22 Student-Level Scale Scores on *SOL* Assessments Correlated with *Stanford 9* Scale Scores

***SOL* Assessment (Spring 1998)**	***Stanford 9* (Spring 1998)**	**Number of Students**	**Pearson Correlation**
Grade 3	Grade 4		
English: Reading & Writing	Total Reading	64,689	.75
Mathematics	Total Mathematics	64,689	.80
Grade 5	Grade 6		
English: Reading/Lit. & Rsrch	Total Reading	61,886	.77
Mathematics	Total Mathematics	61,886	.79
Grade 8	Grade 9		
English: Reading/Lit. & Rsrch	Total Reading	54,881	.76
Mathematics	Total Mathematics	54,881	.82

4. CALIBRATION, EQUATING, AND SCALING PROCEDURES

The IRT model used to develop, calibrate, equate, and scale the Virginia *SOL* assessments was the *Rasch model* (Rasch, 1980) and its polytomous extension, the *Masters Partial Credit model* (PCM) (Masters, 1982). Both of these measurement models have been used for some time to construct test forms, for scaling and equating, and to develop and maintain large item banks.

All test analyses, including item-fit analysis, scaling, equating, diagnosis, and performance prediction were accomplished within this framework. All analyses for the grades 5 and 8, and end-of-course writing tests were based on the Masters Partial Credit model; i.e., multiple-choice items and writing domain scores were combined to form a single scale, and items from different assessment modes and from different test forms were processed simultaneously. The statistical software used to calibrate, scale, and equate the *SOL* assessments included *SAS* (1989), *BIGSTEPS* (Linacre & Wright, 1991), and *TRIAN* (Rentz, 1980).

The technical note following this section outlines the formulation of the Rasch and Partial Credit models in greater detail.

4.1 Equating and Scale Score Derivation Procedures

Equating of operational test forms involved ensuring that all forms in a content area and grade level test (e.g., grade 3 Mathematics) are as equally difficult as possible, both within and across assessment administrations. By equating, students taking one form of a test were neither advantaged nor disadvantaged compared with students taking a different form of a test.

Equating of the *SOL* assessments involved the use of common items on each form of the test. Each test form contained a subset of items that was reproduced on every other test form for the same subject and grade. These items, called *linking items*, served as an anchor for comparison. Each time a new test form is constructed in the future, an attempt will be made to make the new form equal in difficulty to the previous form. This equating was accomplished through statistical procedures using data collected on items during field tests. The data collection design used was the Design IV procedure for common item, non-equivalent groups (Angoff, 1971).

For each test form at a given grade level and content area, the Rasch model was applied in order to obtain parameter estimates for both the unique items on each form, as well as the linking items. The parameter estimates for each form were placed on a common metric by using the Rasch equating constant procedure (Wright & Stone, 1979). This resulted in the item parameters for *all* forms being on the same Rasch ability scale. A consequence of this was that, given an ability estimate θ_n, it was possible to determine scores on different forms that could be considered equivalent.

The final step consisted of obtaining for each raw score point on a form the Rasch ability score or theta corresponding to it. This was done by iteratively solving the expression

$$\eta = \sum_{i=1}^{I} P_{nxi}(\theta_n) \qquad (4.1)$$

where η is the true score associated with student n of ability θ_n, and $P_{nxi}(\theta_n)$ is the probability of a correct response for the PCM for each of the I items and/or task-steps on the form.

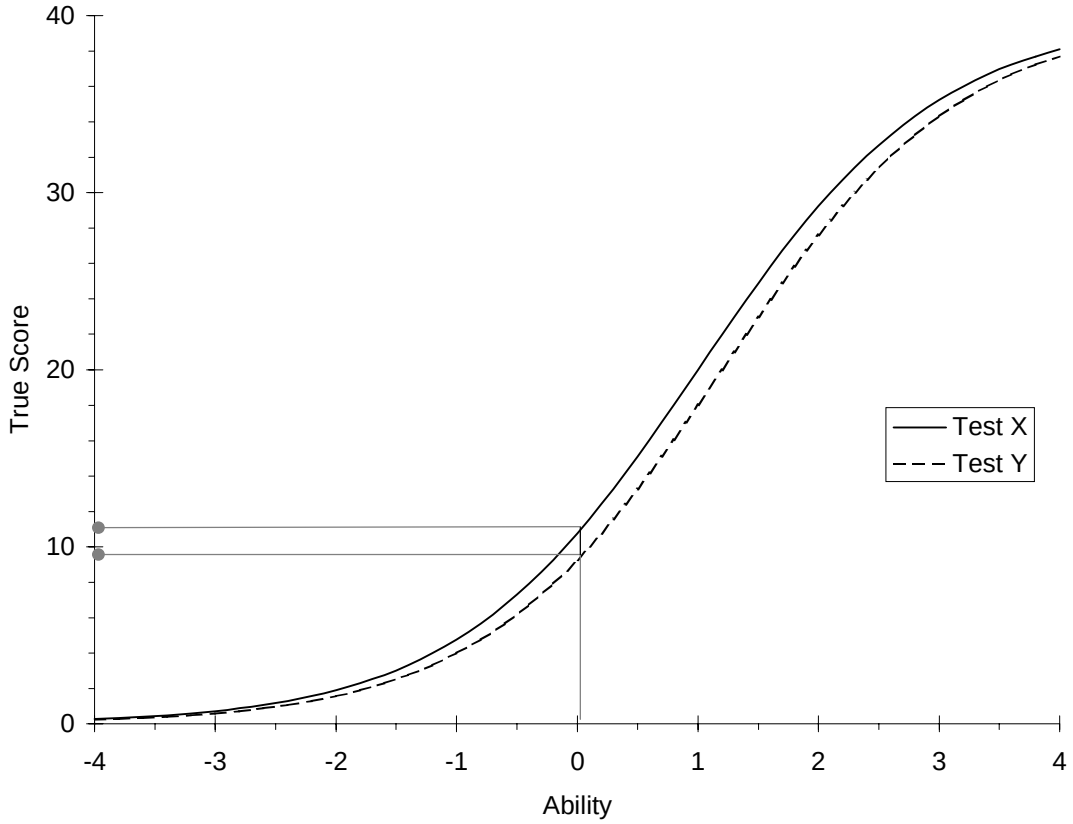

Figure 4.1 True Score Equating

Figure 4.1 illustrates these ideas for two hypothetical test forms, X and Y. In this figure, the true scores on each of the forms are plotted against Rasch ability using Equation 4.1. By drawing a line from the Rasch ability (here shown for an ability of 0) to each of the respective curves, and moving across to the true score scale, one can find the pairs of true scores that are equated to one another. According to Lord and Wingersky (1983), the procedure applied to true scores can safely be transferred to observed scores without any major anomalies in the resulting outcomes.

All post-equating on live test forms was carried out at the total score level, while pre-equating of forms was conducted at the reporting category level. Consequently, as new test forms are developed, they will be of approximate equal difficulty at the reporting category level. Data from these analyses were also used for item review by members of the Content Review Committees.

In order to facilitate the use and interpretation of the *SOL* assessment results, various scale scores were derived for reporting purposes.

Scale Scores for Content Areas

To accomplish the transformation, two levels, d_1 and d_2, were selected on the Rasch ability or theta scale corresponding to standards-referenced criteria. These values were converted to the new scale at easy-to-remember locations, D_1 and D_2. Specifically, $D_1 = 400$ was linked to the cutpoint between *Below Proficient* and *Proficient*, and $D_2 = 500$ was linked with the cut scores

between *Proficient* and *Advanced.* Since d_1 and d_2 were criterion values on the theta scale, and D_1 and D_2 were the values on the new scale, the linear transformation was given by:

$$ScaleScore = \alpha + \gamma \cdot Theta$$

where the slope of the linear transformation is $\alpha = (D_1 d_2 - D_2 d_1)/(d_2 - d_1)$ and the intercept $\gamma = (D_2 - D_1)/(d_2 - d_1)$ (see Wright & Stone, 1979).

This transformation preserved the standards-referenced interpretation of the scale scores by being explicitly linked to the standards-referenced cut scores obtained from the Virginia *SOL* assessment standard setting. In other words, regardless of what form or administration year of the *SOL* assessment, a student would require the same level of ability to obtain a scale score of 400 for proficiency, and a scale score of 500 for advanced. Note that ,while the scale scores can be used for comparisons *within* an *SOL* assessment, they cannot be compared *across* different *SOL* assessment content areas.

It should also be noted that scale scores represent a non-linear transformation of the raw scores from which they were obtained. That is, the distance between scale scores does not remain the same for each change in the raw scores. Typically, for the middle of the scale (around the 350 to 400 range), the increments are smaller than near the top or bottom of the scale. To complete the scale, a scale score of 0 was set to correspond to a raw score of 0, and a scale score of 600 was set to correspond to a perfect raw score.

Scale Scores for Reporting Categories

Scale scores for Reporting Categories in the 1998 *SOL* administration were calculated to provide a norm-referenced interpretation[6].

First, the mean and standard deviation of the theta distribution of each content area was calculated. Next, these values were used to convert each student's Rasch ability or theta to an intermediate scale with a mean of 0 and a standard deviation of 1 by:

$$Z_{98} = (Theta - Mean_{98}) / SD_{98}$$

The final scale for the reporting categories was obtained by converting the intermediate scale to a scale with a mean of 35 and a standard deviation of 5 by:

$$ReportingCategoryScaleScore = 5 \cdot Z_{98} + 35.$$

4.2 Item Bank Construction

The number of test forms to be constructed each year and the need to replace items that would be released to the public necessitated the availability of a large pool of items. The *SOL* item bank was maintained by Harcourt Educational Measurement both in the form of a computer file and a paper copy, making test items readily available to both Harcourt and VDOE staff for reference, test construction, test booklet design, and printing.

[6] In all future *SOL* assessments, scale scores for Reporting Categories will be standards-referenced. These scales will be developed in a process similar to the one used for the Content Area scale scores.

Harcourt Educational Measurement maintains a computerized statistical item bank to store supporting and identification information on each item. The information stored in this item bank includes each item's code number, grade level, content area, *SOL* and reporting category, field test date, test form, and item statistics. The statistical item bank also contains information that resulted from data review meetings. This item statistic information was used during test construction to calculate and adjust for test difficulty, content coverage, and pre-equating test forms and to print individual test statistics as needed.

After the spring 1998 operational administration of the *SOL* assessments, the item bank Rasch scale statistics were re-calibrated using all of the student responses. The re-calibrated scale will serve as the base scale. Standards were set using the 1998 forms as the base year, and future administrations of the tests will be equated to the scales from the base year administration using a common item non-equivalent groups design.

4.3 Summary Tables of the Scaling Results

The raw score to scale score conversions are presented at the end of this section. Tables 4.1 through 4.24 present the results for the Core 1 and Core 2 forms in each grade and content area for the multiple-choice assessments. Tables 4.25 through 4.27 present the conversion tables for the writing assessments for each combination of multiple-choice section and writing prompt.

The results of factor analyses to examine the assumption of unidimensionality underlying the Rasch model are presented in Table 4.28 for the *SOL* multiple-choice assessments and in Table 4.29 for grades 8 and end-of-course of the writing assessments. These results show that the *SOL* assessments are essentially unidimensional.

Table 4.1 Raw Score to Scale Score Conversion: Grade 3 English: Reading & Writing

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	56	0	56
1	130	56	125	56
2	170	40	165	41
3	194	33	190	34
4	212	29	208	30
5	225	26	222	27
6	237	25	235	25
7	247	23	245	24
8	257	22	255	23
9	265	21	264	21
10	273	20	272	21
11	280	20	280	20
12	287	19	287	19
13	293	19	293	19
14	299	18	299	19
15	305	18	306	18
16	312	18	312	18
17	317	18	317	18
18	323	18	323	18
19	328	17	329	18
20	334	17	335	18
21	338	17	340	17
22	344	17	345	17
23	349	17	350	17
24	354	17	356	17
25	360	17	361	18
26	365	17	367	18
27	371	18	372	18
28	376	18	378	18
29	382	18	384	18
30	388	18	389	18
31	394	18	395	19
32	400	19	402	19
33	407	19	409	19
34	413	20	415	20
35	420	20	423	20
36	429	21	431	21
37	437	22	439	22
38	446	23	448	23
39	457	25	459	25
40	468	26	471	26
41	482	29	485	29
42	500	34	503	34
43	524	40	526	40
44	564	56	566	56
45	600	56	600	56

Table 4.2 Raw Score to Scale Score Conversion: Grade 3 Mathematics

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	75	0	75
1	19	75	21	75
2	73	54	75	54
3	105	45	107	45
4	129	40	131	40
5	148	36	150	36
6	163	33	165	33
7	177	31	179	31
8	190	30	191	30
9	201	28	203	28
10	210	27	213	27
11	220	26	223	26
12	229	25	232	25
13	237	25	240	25
14	246	25	248	25
15	254	24	256	24
16	261	24	264	24
17	269	23	271	23
18	275	23	278	23
19	282	22	285	23
20	290	22	292	22
21	296	22	299	22
22	303	22	306	22
23	309	22	312	22
24	316	22	319	22
25	322	22	326	22
26	329	22	332	22
27	336	22	339	22
28	343	22	346	22
29	349	22	353	22
30	356	22	359	22
31	363	23	366	23
32	369	23	374	23
33	377	23	381	23
34	384	24	388	24
35	392	24	396	25
36	400	25	404	25
37	408	25	412	25
38	417	26	421	26
39	426	27	430	27
40	436	28	440	28
41	446	28	450	28
42	457	30	462	30
43	470	31	474	31
44	484	34	488	34
45	500	36	505	36
46	519	40	524	40
47	543	45	548	46
48	575	54	581	54
49	587	76	591	76
50	600	76	600	76

Table 4.3 Raw Score to Scale Score Conversion: Grade 3 History & Social Science

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	54	0	54
1	158	54	160	54
2	199	40	199	39
3	224	34	224	33
4	243	30	242	29
5	258	27	257	27
6	271	25	269	25
7	282	24	280	24
8	292	23	290	23
9	302	21	299	21
10	310	21	308	20
11	318	20	316	20
12	325	19	323	19
13	332	19	330	19
14	339	19	338	19
15	346	18	344	18
16	352	18	350	18
17	358	18	356	18
18	364	18	363	18
19	370	18	369	18
20	376	18	375	18
21	382	18	381	18
22	388	18	387	18
23	394	18	393	18
24	400	18	399	18
25	406	18	406	18
26	413	18	412	18
27	419	18	418	19
28	426	19	425	19
29	432	19	432	19
30	440	20	440	20
31	448	20	447	20
32	455	21	456	21
33	465	23	465	23
34	475	24	475	24
35	486	26	487	26
36	500	28	501	28
37	517	32	518	32
38	540	38	541	38
39	578	53	579	53
40	600	53	600	53

Table 4.4 Raw Score to Scale Score Conversion: Grade 3 Science

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	61	0	61
1	107	61	103	61
2	150	44	147	44
3	177	37	174	37
4	198	32	194	32
5	213	30	210	30
6	227	28	224	28
7	239	26	237	26
8	250	25	247	25
9	260	24	257	24
10	269	23	266	23
11	278	23	275	23
12	287	22	284	22
13	295	22	292	22
14	303	22	300	22
15	310	22	308	22
16	318	21	316	21
17	325	21	323	21
18	333	21	330	21
19	340	21	337	21
20	347	21	344	21
21	354	21	351	21
22	362	21	359	21
23	369	21	366	21
24	377	22	374	21
25	384	22	381	22
26	392	22	389	22
27	400	22	397	22
28	408	23	405	23
29	417	23	414	23
30	426	23	423	23
31	435	25	432	25
32	446	25	443	25
33	457	26	454	26
34	469	28	466	28
35	483	30	480	30
36	500	33	496	33
37	520	37	517	37
38	547	44	544	44
39	592	61	588	61
40	600	61	600	61

Table 4.5 Raw Score to Scale Score Conversion: Grade 5 English: Reading/Literature & Research

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	51	0	52
1	163	51	148	52
2	199	37	186	38
3	222	31	209	31
4	238	27	226	28
5	251	25	240	25
6	262	23	252	24
7	272	22	262	22
8	281	21	271	21
9	289	20	280	21
10	296	19	288	20
11	303	19	295	19
12	310	18	303	19
13	316	18	309	18
14	323	18	316	18
15	328	17	322	18
16	334	17	328	18
17	340	17	334	18
18	345	17	340	17
19	351	17	346	17
20	356	17	352	17
21	362	17	358	17
22	367	17	363	17
23	372	17	369	17
24	378	17	375	17
25	383	17	381	18
26	389	17	387	18
27	394	17	393	18
28	400	18	399	18
29	406	18	406	18
30	413	18	412	19
31	419	19	419	19
32	426	19	427	20
33	433	20	435	20
34	441	21	443	21
35	450	22	452	22
36	460	23	463	24
37	471	25	474	25
38	484	27	488	28
39	500	31	505	31
40	522	37	527	37
41	559	51	565	52
42	600	51	600	52

Table 4.6 Raw Score to Scale Score Conversion: Grade 5 Mathematics

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	54	0	54
1	128	54	128	54
2	166	39	167	39
3	190	32	191	33
4	207	28	208	28
5	221	26	222	26
6	233	24	234	24
7	243	23	245	23
8	253	22	254	22
9	261	21	263	21
10	268	20	271	20
11	276	19	278	19
12	283	19	285	19
13	289	18	292	18
14	295	18	298	18
15	302	17	304	18
16	307	17	310	17
17	313	17	316	17
18	318	17	321	17
19	323	17	327	17
20	328	16	332	17
21	334	16	337	17
22	339	16	343	17
23	344	16	348	16
24	349	16	353	16
25	354	16	358	16
26	359	16	363	16
27	364	16	368	16
28	369	16	374	17
29	374	16	379	17
30	379	16	384	17
31	384	17	389	17
32	389	17	395	17
33	395	17	401	17
34	400	17	406	17
35	406	17	412	18
36	412	18	418	18
37	418	18	425	18
38	424	18	432	19
39	431	19	438	19
40	438	20	446	20
41	446	21	454	21
42	455	22	463	22
43	464	23	473	23
44	474	24	483	25
45	486	26	495	26
46	500	28	510	29
47	517	33	528	33
48	541	39	552	39
49	580	54	592	54
50	600	54	600	54

Table 4.7 Raw Score to Scale Score Conversion: Grade 5 History & Social Science

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	51	0	51
1	175	51	171	51
2	212	37	208	37
3	234	31	231	31
4	250	27	248	27
5	263	25	261	25
6	274	23	272	23
7	284	22	282	22
8	292	21	291	21
9	301	20	299	20
10	308	19	307	19
11	315	19	314	19
12	321	18	321	18
13	327	18	328	18
14	333	17	334	18
15	339	17	340	17
16	345	17	345	17
17	350	17	351	17
18	356	17	357	17
19	361	17	362	17
20	367	17	368	17
21	372	17	373	17
22	377	17	379	17
23	383	17	384	17
24	388	17	390	17
25	394	17	396	17
26	400	17	402	18
27	406	18	408	18
28	413	18	415	18
29	419	19	421	19
30	426	19	428	19
31	433	20	435	20
32	441	21	443	21
33	450	22	452	22
34	460	23	462	23
35	471	25	473	25
36	484	27	486	27
37	500	31	503	31
38	523	37	525	37
39	559	51	561	51
40	600	51	600	51

Table 4.8 Raw Score to Scale Score Conversion: Grade 5 Science

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	50	0	50
1	165	50	179	50
2	201	37	215	36
3	224	31	237	30
4	241	27	253	26
5	255	25	266	24
6	267	23	277	22
7	277	22	287	21
8	286	21	295	20
9	295	20	304	20
10	302	19	311	19
11	310	19	318	18
12	317	18	325	18
13	323	18	331	18
14	330	18	337	17
15	336	17	343	17
16	342	17	349	17
17	348	17	354	17
18	354	17	360	17
19	360	17	365	17
20	365	17	371	17
21	371	17	376	17
22	377	17	382	17
23	382	17	387	17
24	388	17	393	17
25	394	17	399	17
26	400	17	405	17
27	406	18	411	18
28	413	18	417	18
29	419	18	423	18
30	426	19	431	19
31	434	20	438	20
32	441	20	446	20
33	450	21	454	21
34	460	22	464	22
35	471	24	475	24
36	484	26	488	26
37	500	30	504	30
38	522	36	526	36
39	558	50	561	50
40	600	50	600	50

Table 4.9 Raw Score to Scale Score Conversion: Grade 5 Computer/Technology

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	48	0	48
1	214	48	215	48
2	250	35	250	35
3	271	29	271	29
4	287	26	288	26
5	300	24	301	24
6	312	22	312	22
7	322	21	322	21
8	331	21	331	20
9	340	20	340	20
10	348	19	348	19
11	356	19	355	19
12	364	19	362	18
13	371	19	370	18
14	379	18	376	18
15	386	18	383	18
16	393	18	390	18
17	400	19	397	18
18	407	19	404	18
19	415	19	411	19
20	423	19	419	19
21	431	20	426	19
22	440	21	434	20
23	449	21	443	21
24	459	22	453	22
25	471	24	464	23
26	484	26	477	26
27	500	29	493	29
28	521	35	514	35
29	557	48	549	48
30	600	48	600	48

Table 4.10 Raw Score to Scale Score Conversion: Grade 8 English: Reading/Literature & Research

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	65	0	67
1	98	65	99	67
2	145	47	148	48
3	174	40	178	40
4	196	35	201	35
5	213	32	219	32
6	227	29	233	30
7	240	28	246	28
8	251	26	258	27
9	262	26	269	26
10	272	24	278	24
11	281	24	288	24
12	290	23	296	23
13	297	23	305	22
14	306	22	312	22
15	313	22	320	22
16	321	22	327	22
17	328	22	334	21
18	335	21	341	21
19	342	21	348	21
20	349	21	355	21
21	356	21	362	21
22	363	21	369	21
23	371	21	376	21
24	378	21	382	21
25	385	22	389	21
26	392	22	396	21
27	400	22	403	22
28	407	22	411	22
29	415	23	419	22
30	424	23	426	23
31	432	24	435	23
32	441	24	444	24
33	451	26	453	25
34	462	26	463	26
35	473	28	474	28
36	486	29	487	29
37	500	31	500	31
38	517	35	517	35
39	538	39	538	39
40	567	47	566	47
41	596	65	594	65
42	600	65	600	65

Table 4.11 Raw Score to Scale Score Conversion: Grade 8 Mathematics

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	50	0	50
1	160	50	158	50
2	196	36	194	35
3	218	30	215	30
4	233	26	231	26
5	246	24	244	24
6	257	22	255	22
7	266	20	264	20
8	274	19	272	19
9	281	18	279	18
10	288	17	286	18
11	294	17	292	17
12	300	17	298	17
13	305	16	304	16
14	310	16	309	16
15	315	15	314	16
16	320	15	319	15
17	324	15	323	15
18	329	15	327	15
19	333	14	332	15
20	337	14	336	14
21	341	14	340	14
22	345	14	344	14
23	349	14	348	14
24	352	14	352	14
25	356	14	356	14
26	360	14	359	14
27	364	13	363	14
28	367	13	367	14
29	371	13	371	14
30	375	13	374	14
31	378	13	378	14
32	382	13	382	14
33	385	13	386	14
34	389	13	389	14
35	393	14	393	14
36	397	14	397	14
37	400	14	401	14
38	404	14	404	14
39	408	14	408	14
40	412	14	412	14
41	416	14	416	14
42	420	14	421	15
43	424	15	425	15
44	429	15	429	15
45	433	15	434	15
46	438	16	439	16
47	443	16	444	16
48	448	16	449	17
49	454	17	455	17
50	460	17	461	17
51	466	18	467	18
52	473	19	474	19
53	481	20	482	20
54	490	21	491	21
55	500	23	501	23
56	512	25	513	26
57	527	29	528	29
58	548	35	550	35
59	583	49	584	49
60	600	49	600	49

Table 4.12 Raw Score to Scale Score Conversion: Grade 8 History & Social Science

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	60	0	60
1	104	60	108	60
2	147	43	151	43
3	174	36	177	36
4	193	32	197	32
5	209	29	212	29
6	222	27	225	26
7	233	25	236	25
8	243	24	246	23
9	253	23	255	23
10	261	22	263	22
11	269	21	272	21
12	277	20	279	20
13	284	20	286	20
14	290	20	293	19
15	296	19	299	19
16	303	19	305	19
17	309	19	311	19
18	315	19	317	18
19	321	19	322	18
20	327	18	328	18
21	332	18	334	18
22	338	18	339	18
23	343	18	345	18
24	349	18	350	18
25	354	18	355	18
26	360	18	360	18
27	365	18	366	18
28	371	18	372	18
29	377	18	377	18
30	382	18	383	18
31	388	19	388	18
32	394	19	394	18
33	400	19	400	19
34	406	19	405	19
35	412	19	412	19
36	419	20	418	19
37	426	20	425	20
38	433	20	432	20
39	440	21	439	21
40	448	22	447	22
41	457	23	455	23
42	465	24	465	23
43	476	25	475	25
44	487	26	486	26
45	500	29	498	29
46	515	32	513	31
47	535	36	532	36
48	560	43	558	43
49	585	60	584	60
50	600	60	600	60

Table 4.13 Raw Score to Scale Score Conversion: Grade 8 Science

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	50	0	50
1	180	50	177	50
2	215	36	213	36
3	237	29	235	30
4	252	26	251	26
5	265	24	264	24
6	275	22	274	22
7	284	21	284	21
8	293	20	293	20
9	300	19	300	19
10	307	18	307	18
11	313	18	314	18
12	320	17	320	17
13	325	17	326	17
14	331	16	331	16
15	336	16	337	16
16	341	16	342	16
17	346	16	347	16
18	350	15	352	15
19	355	15	356	15
20	360	15	361	15
21	365	15	366	15
22	369	15	370	15
23	374	15	375	15
24	378	15	379	15
25	382	15	384	15
26	387	15	388	15
27	391	15	393	15
28	396	15	397	15
29	400	15	401	15
30	405	15	406	15
31	409	15	411	15
32	414	15	416	15
33	419	16	420	16
34	424	16	425	16
35	429	16	430	16
36	434	16	436	16
37	440	17	441	17
38	445	17	447	17
39	451	18	453	18
40	458	18	459	18
41	465	19	466	19
42	472	20	474	20
43	480	21	482	21
44	489	22	492	22
45	500	24	502	24
46	512	26	515	26
47	528	29	530	30
48	550	36	552	36
49	585	50	588	50
50	600	50	600	50

Table 4.14 Raw Score to Scale Score Conversion: Grade 8 Computer/Technology

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	56	0	56
1	134	56	144	56
2	175	41	185	41
3	200	34	209	34
4	218	30	228	30
5	234	27	243	27
6	246	26	255	25
7	258	24	266	24
8	268	23	276	23
9	277	23	286	22
10	286	21	295	21
11	294	21	302	21
12	302	21	310	20
13	310	20	318	20
14	317	20	325	20
15	324	20	332	20
16	331	20	338	19
17	338	19	346	19
18	345	19	352	19
19	352	19	359	19
20	358	19	365	19
21	365	19	372	19
22	372	19	379	19
23	379	20	385	19
24	386	20	392	19
25	393	20	399	20
26	400	20	406	20
27	408	20	414	20
28	415	21	421	21
29	423	21	429	21
30	432	22	437	21
31	441	23	446	23
32	450	23	455	23
33	460	24	466	24
34	472	26	477	26
35	485	27	490	27
36	500	30	505	30
37	519	34	524	34
38	543	41	549	41
39	584	56	590	57
40	600	56	600	57

Table 4.15 Raw Score to Scale Score Conversion: End-of-Course English: Reading/Literature & Research

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	55	0	55
1	167	55	178	55
2	207	40	217	39
3	231	33	241	33
4	249	29	259	29
5	263	26	273	26
6	275	25	285	25
7	286	23	295	23
8	295	22	305	22
9	304	21	313	21
10	312	20	321	20
11	320	20	329	20
12	327	19	336	19
13	334	19	343	19
14	340	19	349	19
15	347	18	355	18
16	353	18	362	18
17	359	18	368	18
18	365	18	374	18
19	371	18	379	18
20	376	18	385	18
21	382	18	391	18
22	388	18	397	18
23	394	18	402	18
24	400	18	408	18
25	406	18	414	18
26	412	18	420	18
27	418	18	426	18
28	424	19	432	19
29	430	19	439	19
30	437	19	446	19
31	444	20	453	20
32	452	20	460	20
33	460	21	468	21
34	468	22	476	22
35	478	23	486	23
36	488	24	496	24
37	500	26	508	26
38	514	29	522	29
39	531	33	539	33
40	555	39	563	39
41	594	55	587	55
42	600	55	600	55

Table 4.16 Raw Score to Scale Score Conversion: End-of-Course US History

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	61	0	61
1	94	61	97	61
2	138	44	141	44
3	164	37	167	36
4	184	32	186	32
5	199	29	201	29
6	211	27	215	27
7	223	25	226	25
8	233	23	235	23
9	242	23	244	23
10	249	22	253	22
11	257	21	260	21
12	264	20	267	20
13	271	20	274	20
14	277	19	280	19
15	283	19	286	19
16	289	19	291	19
17	295	18	297	18
18	299	18	302	17
19	305	17	307	17
20	310	17	312	17
21	314	17	317	17
22	319	17	322	17
23	324	17	327	17
24	329	17	331	16
25	333	16	336	16
26	338	16	340	16
27	342	16	344	16
28	346	16	349	16
29	351	16	353	16
30	355	16	357	16
31	360	16	362	16
32	364	16	366	16
33	369	16	371	16
34	373	16	375	16
35	377	16	379	16
36	381	16	383	16
37	386	16	388	16
38	390	17	392	16
39	395	17	397	17
40	400	17	401	17
41	405	17	406	17
42	410	17	411	17
43	414	17	417	17
44	420	18	421	18
45	425	18	427	18
46	431	19	432	19
47	437	19	438	19
48	443	19	444	19
49	449	20	450	20
50	456	20	457	20
51	463	22	465	21
52	471	22	472	22
53	480	23	481	23
54	489	25	490	25
55	500	27	501	27
56	513	28	513	28
57	527	31	528	31
58	546	36	547	36
59	572	43	573	43
60	598	61	599	61
61	600	61	600	61

Table 4.17 Raw Score to Scale Score Conversion: End-of-Course World History to 1000 A.D./World Geography

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	45	0	45
1	198	45	197	45
2	230	33	230	33
3	249	27	249	27
4	263	24	263	24
5	274	22	275	22
6	284	20	284	20
7	292	18	292	19
8	299	17	299	17
9	305	17	306	17
10	312	16	312	16
11	317	16	318	16
12	322	15	323	15
13	327	15	328	15
14	332	14	333	14
15	336	14	337	14
16	340	14	342	14
17	344	13	346	13
18	348	13	350	13
19	352	13	354	13
20	356	13	357	13
21	360	13	361	13
22	363	13	365	13
23	367	13	369	13
24	370	13	372	13
25	374	13	375	13
26	377	12	379	13
27	380	12	382	13
28	384	12	386	12
29	387	12	389	12
30	390	12	392	12
31	394	12	395	12
32	397	12	399	12
33	400	12	402	12
34	404	12	406	12
35	407	12	409	13
36	410	13	412	13
37	414	13	416	13
38	417	13	419	13
39	421	13	423	13
40	424	13	426	13
41	428	13	430	13
42	431	13	434	13
43	435	13	438	13
44	439	13	442	13
45	443	14	446	14
46	448	14	450	14
47	452	14	455	14
48	457	15	459	15
49	461	15	464	15
50	467	16	469	16
51	472	16	475	16
52	478	17	481	17
53	485	17	487	17
54	492	18	495	18
55	500	20	503	20
56	509	22	512	22
57	521	24	523	24
58	535	27	537	27
59	554	33	556	33
60	586	45	589	45
61	600	45	600	45

Table 4.18 Raw Score to Scale Score Conversion: End-of-Course World History from 1000 A.D./World Geography

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	49	0	50
1	175	49	174	50
2	210	35	209	35
3	231	29	230	30
4	246	26	246	26
5	258	23	259	23
6	268	21	269	21
7	277	20	278	20
8	285	19	285	19
9	292	18	292	18
10	298	17	299	17
11	304	17	305	17
12	310	16	311	16
13	315	16	316	16
14	319	15	320	15
15	324	15	325	15
16	329	15	330	15
17	333	14	334	15
18	337	14	338	14
19	341	14	342	14
20	345	14	346	14
21	349	14	350	14
22	352	13	353	13
23	356	13	357	13
24	359	13	3610	13
25	363	13	364	13
26	367	13	367	13
27	370	13	371	13
28	373	13	374	13
29	376	13	378	13
30	380	13	381	13
31	383	13	384	13
32	386	13	388	13
33	390	13	391	13
34	393	13	394	13
35	397	13	398	13
36	400	13	401	13
37	403	13	404	13
38	407	13	408	13
39	410	13	411	13
40	414	13	415	13
41	417	13	418	13
42	421	14	422	14
43	425	14	425	14
44	428	14	430	14
45	433	14	433	14
46	436	14	437	14
47	441	15	441	15
48	445	15	446	15
49	450	15	450	15
50	454	16	455	16
51	460	16	461	16
52	465	17	466	17
53	471	17	472	17
54	477	18	478	18
55	484	19	485	19
56	491	20	492	20
57	500	21	501	21
58	510	23	511	23
59	522	25	523	25
60	537	29	538	29
61	558	35	558	35
62	592	49	593	49
63	600	49	600	49

Table 4.19 Raw Score to Scale Score Conversion: End-of-Course Earth Science

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	52	0	52
1	157	52	157	52
2	195	38	195	37
3	218	31	217	31
4	235	28	234	28
5	249	25	247	25
6	260	24	258	23
7	270	22	268	22
8	279	21	277	21
9	287	20	285	20
10	295	19	292	19
11	302	18	299	18
12	308	18	305	18
13	314	18	312	17
14	321	17	317	17
15	326	17	323	17
16	332	17	329	16
17	337	16	334	16
18	343	16	339	16
19	348	16	344	16
20	352	16	349	16
21	357	16	354	16
22	362	16	358	15
23	367	16	363	15
24	372	15	368	15
25	376	15	372	15
26	381	15	377	15
27	386	15	382	15
28	391	16	387	16
29	395	16	392	16
30	400	16	396	16
31	405	16	401	16
32	410	16	406	16
33	415	16	412	16
34	421	16	417	16
35	426	17	422	17
36	431	17	428	17
37	437	17	434	17
38	443	18	440	18
39	449	18	447	18
40	456	19	453	19
41	463	19	460	20
42	471	21	469	21
43	479	22	477	22
44	489	23	488	23
45	500	25	499	25
46	513	27	512	28
47	530	31	529	31
48	552	37	551	37
49	589	52	589	52
50	600	52	600	52

Table 4.20 Raw Score to Scale Score Conversion: End-of-Course Biology

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	44	0	44
1	215	44	215	44
2	246	32	246	32
3	265	26	265	26
4	279	23	279	23
5	290	21	290	21
6	300	19	299	19
7	308	18	307	18
8	316	17	314	17
9	322	17	320	16
10	328	16	327	16
11	334	16	332	15
12	339	15	337	15
13	345	15	342	14
14	350	14	347	14
15	354	14	352	14
16	359	14	356	14
17	363	14	360	13
18	368	13	364	13
19	372	13	368	13
20	376	13	372	13
21	380	13	376	13
22	384	13	380	13
23	388	13	384	13
24	392	13	388	13
25	396	13	391	13
26	400	13	395	13
27	404	13	399	13
28	408	13	403	13
29	412	13	407	13
30	416	13	411	13
31	420	13	415	13
32	424	13	419	13
33	429	14	423	13
34	433	14	427	14
35	438	14	431	14
36	442	14	436	14
37	447	15	441	14
38	452	15	446	15
39	457	16	451	15
40	463	16	456	16
41	469	16	462	16
42	476	17	469	17
43	483	18	476	18
44	491	19	484	19
45	500	21	493	21
46	511	23	504	23
47	525	26	518	26
48	544	32	536	32
49	575	44	568	44
50	600	44	600	44

Table 4.21 Raw Score to Scale Score Conversion: End-of-Course Chemistry

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	45	0	45
1	203	45	203	45
2	235	33	235	32
3	255	27	254	27
4	270	24	269	24
5	281	22	280	22
6	291	20	290	20
7	300	19	298	19
8	308	18	306	18
9	315	17	313	17
10	321	17	319	16
11	327	16	325	16
12	333	15	331	15
13	338	15	336	15
14	343	15	341	15
15	348	15	346	15
16	353	15	350	14
17	358	14	355	14
18	362	14	360	14
19	366	14	364	14
20	371	14	368	14
21	375	14	372	14
22	379	14	377	14
23	383	14	381	13
24	388	14	385	13
25	392	14	389	13
26	396	14	393	13
27	400	14	397	13
28	404	14	401	14
29	408	14	405	14
30	412	14	409	14
31	417	14	413	14
32	421	14	418	14
33	426	14	422	14
34	430	14	427	14
35	435	15	431	15
36	439	15	436	15
37	445	15	441	15
38	450	15	447	15
39	455	16	452	16
40	461	16	458	16
41	468	17	464	17
42	474	18	471	18
43	482	19	478	19
44	490	20	487	20
45	500	22	497	22
46	512	23	508	23
47	526	27	522	27
48	545	32	542	32
49	577	45	573	45
50	600	45	600	45

Table 4.22 Raw Score to Scale Score Conversion: End-of-Course Algebra I

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	45	0	45
1	209	45	209	45
2	241	32	241	32
3	260	27	260	27
4	275	24	275	24
5	286	21	286	21
6	296	20	295	20
7	304	19	303	19
8	311	18	311	18
9	318	17	318	17
10	324	16	324	16
11	330	16	330	16
12	335	16	335	15
13	340	15	340	15
14	345	15	345	15
15	350	14	350	14
16	355	14	354	14
17	359	14	359	14
18	364	14	363	14
19	368	14	367	14
20	372	13	371	13
21	376	13	375	13
22	380	13	379	13
23	384	13	383	13
24	388	13	387	13
25	392	13	391	13
26	396	13	395	13
27	400	13	399	13
28	404	13	403	13
29	408	13	407	13
30	412	14	411	13
31	416	14	415	14
32	420	14	419	14
33	425	14	423	14
34	429	14	428	14
35	434	15	432	14
36	439	15	437	15
37	444	15	442	15
38	449	16	447	15
39	455	16	452	16
40	461	16	458	16
41	467	17	464	17
42	474	18	471	18
43	482	19	478	19
44	490	20	487	20
45	500	22	496	21
46	512	24	507	24
47	526	27	521	27
48	546	33	541	32
49	579	45	573	45
50	600	45	600	45

Table 4.23 Raw Score to Scale Score Conversion: End-of-Course Geometry

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	50	0	50
1	174	50	172	50
2	210	36	208	36
3	232	30	230	30
4	248	27	247	27
5	261	24	260	24
6	273	23	271	22
7	283	21	281	21
8	291	20	289	20
9	300	19	297	19
10	307	19	305	19
11	314	18	312	18
12	320	18	318	17
13	327	17	324	17
14	333	17	330	17
15	339	16	336	16
16	344	16	342	16
17	349	16	347	16
18	355	16	353	16
19	360	16	358	16
20	365	15	363	15
21	370	15	368	15
22	375	15	373	15
23	380	15	378	15
24	385	15	383	15
25	390	15	388	15
26	395	15	393	15
27	400	15	398	16
28	405	16	403	16
29	410	16	409	16
30	415	16	414	16
31	421	16	419	16
32	427	17	425	17
33	432	17	431	17
34	439	17	438	18
35	445	18	444	18
36	452	19	451	19
37	459	19	459	20
38	468	20	467	21
39	477	22	476	22
40	487	24	487	24
41	500	26	500	26
42	515	29	515	29
43	537	35	537	35
44	572	49	572	49
45	600	49	600	49

Table 4.24 Raw Score to Scale Score Conversion: End-of-Course Algebra II

	Core 1		Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error
0	0	56	0	56
1	141	56	141	56
2	182	40	181	40
3	206	34	205	33
4	224	29	223	29
5	238	27	237	27
6	250	25	249	25
7	261	23	259	23
8	270	22	269	22
9	278	21	277	21
10	286	20	285	20
11	293	20	292	19
12	300	19	299	19
13	307	19	305	19
14	313	18	311	18
15	319	18	317	18
16	324	18	322	18
17	330	17	328	17
18	335	17	333	17
19	340	17	338	17
20	345	17	343	17
21	350	17	348	17
22	355	17	353	17
23	360	17	358	17
24	365	17	363	17
25	370	17	368	17
26	375	17	373	17
27	380	17	378	17
28	385	17	383	17
29	390	17	388	17
30	395	17	392	17
31	400	17	397	17
32	405	17	402	17
33	410	17	408	17
34	415	18	413	18
35	422	18	419	18
36	427	18	425	18
37	433	18	431	18
38	439	19	437	19
39	446	19	443	19
40	453	20	450	20
41	461	21	458	21
42	469	22	467	22
43	478	23	475	23
44	488	24	486	24
45	500	27	497	27
46	514	29	511	29
47	531	33	529	33
48	555	40	553	40
49	595	56	592	56
50	600	56	600	56

Table 4.25 Raw Score to Scale Score Conversion: Grade 5 Writing (by Writing Prompt /Multiple-Choice Combination)

	Core 1/Core 1		Core 1/Core 2		Core 2/Core 1		Core 2/Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error	Scale Score	Standard Error	Scale Score	Standard Error
0	0	46	0	46	0	45	0	46
1	30	46	37	46	33	45	39	46
2	44	46	61	46	57	45	63	46
3	78	46	85	46	81	45	87	46
4	102	46	109	46	105	45	111	46
5	126	46	133	46	129	45	135	46
6	150	46	157	46	153	45	159	46
7	174	46	181	46	177	45	181	46
8	205	32	212	33	207	32	214	32
9	222	26	231	27	225	26	232	27
10	235	23	244	23	238	23	246	23
11	246	22	255	22	248	21	256	21
12	255	20	264	20	257	20	265	20
13	264	20	272	19	265	19	273	19
14	272	19	280	19	273	19	280	18
15	279	19	287	18	281	19	287	18
16	287	19	294	18	288	19	294	18
17	295	19	301	18	296	19	301	18
18	302	19	308	18	303	18	308	18
19	309	19	315	18	310	18	315	18
20	317	18	322	18	318	18	321	18
21	324	18	329	18	325	18	328	18
22	331	18	336	18	332	18	335	18
23	338	18	343	18	339	18	342	18
24	345	18	350	18	346	18	349	18
25	352	18	357	18	353	18	356	18
26	359	18	363	18	359	18	363	18
27	365	18	370	18	366	18	370	18
28	372	18	377	18	372	18	377	18
29	378	18	384	18	379	18	384	18
30	385	18	391	18	386	18	391	18
31	392	18	399	19	393	18	399	19
32	400	19	406	19	400	19	406	19
33	408	19	414	19	408	19	414	20
34	416	20	422	20	416	20	423	20
35	425	21	431	21	425	21	431	21
36	434	22	440	22	435	22	441	22
37	445	23	451	23	445	23	453	23
38	456	24	463	24	456	24	465	25
39	469	25	476	26	469	25	480	28
40	484	27	492	28	484	26	498	30
41	500	29	509	30	500	29	518	31
42	520	33	530	34	519	33	541	35
43	551	45	562	45	550	45	574	46
44	600	45	600	45	600	45	600	46

Table 4.26 Raw Score to Scale Score Conversion: Grade 8 Writing (by Writing Prompt/Multiple-Choice Combination)

	Core 1/Core 1		Core 1/Core 2		Core 2/Core 1		Core 2/Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error	Scale Score	Standard Error	Scale Score	Standard Error
0	0	38	0	38	000	38	0	38
1	50	38	56	38	054	38	52	38
2	77	38	83	38	081	38	79	38
3	104	38	110	38	108	38	106	38
4	131	38	137	38	135	38	133	38
5	158	38	164	38	162	38	160	38
6	185	38	191	38	189	38	187	38
7	212	38	218	38	216	38	214	38
8	238	27	244	26	242	27	240	27
9	253	23	258	22	257	23	255	22
10	265	20	269	19	268	20	266	20
11	274	19	278	18	278	18	275	18
12	283	18	286	17	286	18	284	18
13	291	18	294	17	293	17	291	17
14	298	17	301	16	300	16	298	17
15	306	17	308	16	307	16	305	16
16	313	17	314	16	314	16	312	16
17	320	16	321	16	320	16	319	16
18	327	16	328	16	327	16	326	16
19	333	16	334	16	333	16	333	16
20	340	16	340	16	339	16	339	16
21	346	16	347	15	345	15	346	16
22	353	15	353	15	351	15	352	16
23	358	15	358	15	357	15	358	15
24	364	15	364	15	363	15	363	15
25	370	15	370	15	368	15	370	15
26	376	15	375	15	374	15	375	15
27	382	15	381	15	380	15	381	15
28	388	15	388	15	386	15	387	15
29	393	16	393	16	392	16	393	16
30	400	16	400	16	398	16	400	16
31	406	16	407	16	405	16	406	16
32	413	17	414	17	412	17	413	17
33	421	17	421	18	420	18	421	18
34	428	18	429	18	428	18	429	18
35	437	18	438	19	437	19	437	19
36	446	19	447	19	446	19	447	19
37	455	19	457	2	456	20	456	20
38	465	20	467	2	466	21	467	21
39	475	21	478	21	477	21	478	21
40	487	22	490	22	490	23	491	23
41	500	24	504	24	504	24	505	25
42	517	28	520	28	521	28	523	28
43	544	38	547	38	548	38	550	39
44	600	38	600	38	600	38	600	39

Table 4.27 Raw Score to Scale Score Conversion: End-of-Course Writing (by Writing Prompt /Multiple-Choice Combination)

	Core 1/Core 1		Core 1/Core 2		Core 2/Core 1		Core 2/Core 2	
Raw Score	Scale Score	Standard Error	Scale Score	Standard Error	Scale Score	Standard Error	Scale Score	Standard Error
0	0	45	0	45	000	46	0	45
1	30	45	34	45	025	46	32	45
2	54	45	58	45	049	46	56	45
3	78	45	82	45	073	46	80	45
4	102	45	106	45	097	46	104	45
5	126	45	130	45	121	46	128	45
6	150	45	154	45	145	46	152	45
7	174	45	178	45	169	46	176	45
8	206	33	210	32	201	33	208	33
9	225	27	229	26	221	27	228	27
10	239	24	242	23	235	24	242	24
11	250	21	253	21	246	21	253	21
12	260	20	263	20	255	20	262	20
13	268	19	271	18	263	18	270	18
14	275	18	278	17	271	17	277	17
15	283	17	284	17	277	17	284	17
16	289	17	291	17	283	17	290	16
17	295	16	296	16	289	16	296	16
18	301	16	302	16	295	16	301	15
19	306	16	308	16	300	16	306	15
20	312	16	313	15	306	15	311	15
21	317	15	318	15	311	15	316	15
22	322	15	323	15	316	15	321	15
23	328	15	328	15	321	15	325	14
24	333	15	333	15	326	15	330	14
25	338	15	338	15	331	15	335	14
26	343	15	343	15	336	15	339	14
27	348	15	348	15	342	15	344	14
28	353	15	352	15	346	15	349	14
29	358	15	357	15	351	15	354	15
30	363	15	362	15	357	15	358	15
31	368	15	367	15	362	15	363	15
32	373	15	372	15	367	15	368	15
33	379	15	377	15	372	15	373	15
34	384	15	382	15	377	16	378	15
35	389	16	387	15	383	16	383	15
36	395	16	392	15	388	16	389	16
37	400	16	397	16	394	16	394	16
38	406	16	403	16	400	16	400	16
39	412	17	408	16	406	17	406	17
40	418	17	414	16	412	17	412	17
41	425	17	420	17	419	17	418	17
42	431	18	427	17	425	18	425	17
43	438	18	433	17	433	18	432	18
44	446	19	441	18	441	19	440	19
45	455	20	448	19	450	20	448	20
46	464	21	456	20	459	21	457	21
47	475	22	465	21	469	22	467	21
48	487	24	475	21	481	24	478	23
49	500	25	486	23	494	25	491	25
50	515	27	499	25	510	27	506	28
51	533	30	514	27	528	30	526	33
52	556	34	533	32	550	34	556	42
53	591	46	565	44	585	46	600	63
54	600	46	600	44	600	46	600	63

Table 4.28 Factor Analyses for *SOL* Multiple-Choice Assessments: Proportion of Variability Explained by First Factor

	Proportion of Variability Explained	
Standards of Learning Assessment	**Core 1**	**Core 2**
Grade 3		
English: Reading & Writing	.995	.950
Mathematics	.963	.910
History	>.999	>.999
Science	>.999	.984
Grade 5		
English: Literature & Research	>.999	.950
Mathematics	.867	.844
History	>.999	>.999
Science	>.999	>.999
Computer/Technology	>.999	>.999
Grade 8		
English: Literature & Research	>.999	.968
Mathematics	.964	.858
History	>.999	.914
Science	>.999	.927
Computer/Technology	.988	.932
High school		
English: Literature & Research	.958	.931
Algebra I	.873	.781
Algebra II	.881	.769
Geometry	.980	.876
US History	.978	.894
Wrld Hist to 1000A.D./Wrld Geog	.962	.812
Wrld Hist from 1000A.D./Wrld Geog	.967	.732
Biology	>.999	.960
Earth Science	.979	.900
Chemistry	.974	.892

Table 4.29 Factor Analyses for *SOL* Writing Assessments: Proportion of Variability Explained by First Factor

	Writing Assessment Configuration		
Grade	**Prompt**	**MC**	**Proportion of Variability Explained**
Grade 8	Core 1	Core 1	.985
	Core 1	Core 2	.985
	Core 2	Core 1	.979
	Core 2	Core 2	.872
End-of-Course	Core 1	Core 1	.879
	Core 1	Core 2	.848
	Core 2	Core 1	.869
	Core 2	Core 2	.534

TECHNICAL NOTE: THE RASCH AND PARTIAL CREDIT IRT MODELS

The most basic expression of the Rasch model is in the Item Characteristic Curves (ICC). Item Characteristic Curves are a function of the probability of a correct response to an item at a specified ability level. The probability of a correct response is bounded by 1 (certainty of a correct response) and 0 (certainty of an incorrect response). The ability scale is, in theory, unbounded and can range from -4 to +4. In practice, the ability scale ranges from -3 to +3 logits for heterogeneous ability groups. A logit (natural log odds of a correct response) of zero typically represents "average" ability.

In Figure 1, a person whose ability falls at -1 on the ability (horizontal) scale has a probability of roughly 20% of answering the item correctly. Another way of expressing this is that if we have a group of 100 people, all of whom have an ability of -1, we would expect about 20% of them to answer the item correctly. Similarly, a person whose ability was at +1 would have about a 70% chance of getting the item right. Thus, a person whose ability is above average is more likely to answer the item correctly than is one whose ability is below average. This makes intuitive sense and is the basic formulation of Rasch measurement for test items having only 2 possible categories (i.e., wrong or right).

To extend the formulation, consider that the Item Characteristic Curve shown here represents the Rasch expression that relates a person's ability to the probability of a correct response to a given item. One might ask what sort of curve would represent the other possible condition, that of answering the item incorrectly. Intuitively, it would seem that if one has a probability of 70% of getting the answer right at an ability level of 1, then the probability of getting it wrong is 30%; at -1 on the ability scale, the probability of answering incorrectly is 80%. Thus, the less ability one has, the more likely he or she is to answer a test item incorrectly. This relationship is depicted in Figure 2.

The key step in the formulation, and the point at which the Rasch dichotomous model merges with the Partial Credit model, requires us to posit an additional response category. Suppose that, rather than scoring items as completely wrong or completely right, we add a category representing answers that, though not totally correct, are still clearly not totally incorrect. These relationships are shown in Figure 3.

The left-most curve in Figure 3 represents the distribution of ability for all people getting a score of "0" (completely incorrect) on the item. Those of very low ability (e.g., -3 to -2) are very likely to be in this category and, in fact, are more likely to be in this category than the other two. Those receiving a "1" tend to fall in the middle range of abilities (the middle curve). The final, right-most curve represents the distribution of abilities for those receiving scores of "2" (completely correct). Very high ability people are clearly more likely to be in this category than in any other, but there are still some of average and low ability who can get full credit for the item.

Although the actual computations are quite complex, the points at which lines cross each other have a similar interpretation as for the dichotomous case. Consider the point at which the category 1 line crosses the category 2 line. For abilities to the left of (or less than) this point, the probability is greatest for a category 1 response. To the right, (or above) this point, and up to the point at which the lines cross for categories 2 and 3, the most likely response is category 2.

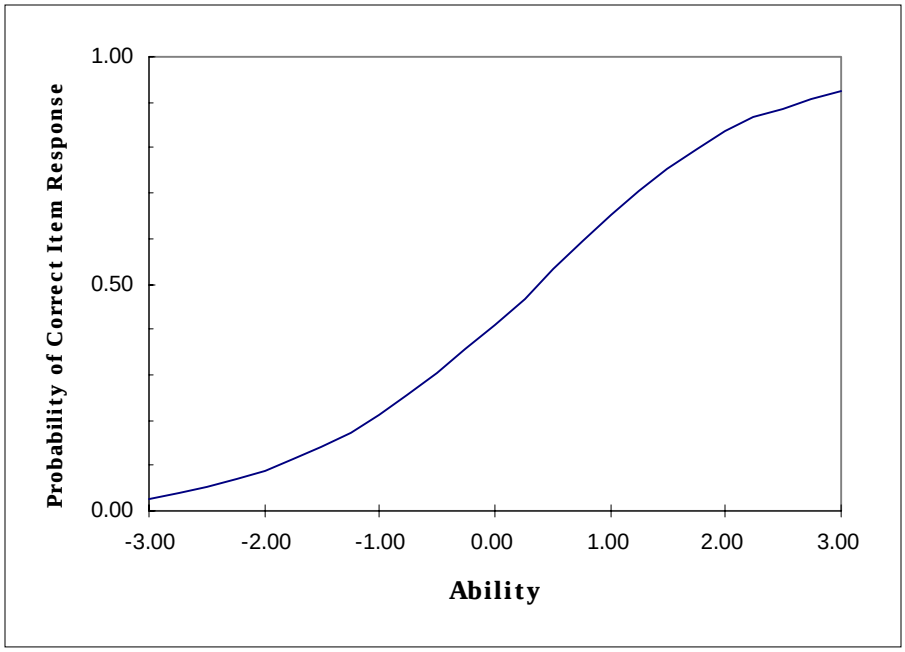

Figure 1 Sample item characteristic curve

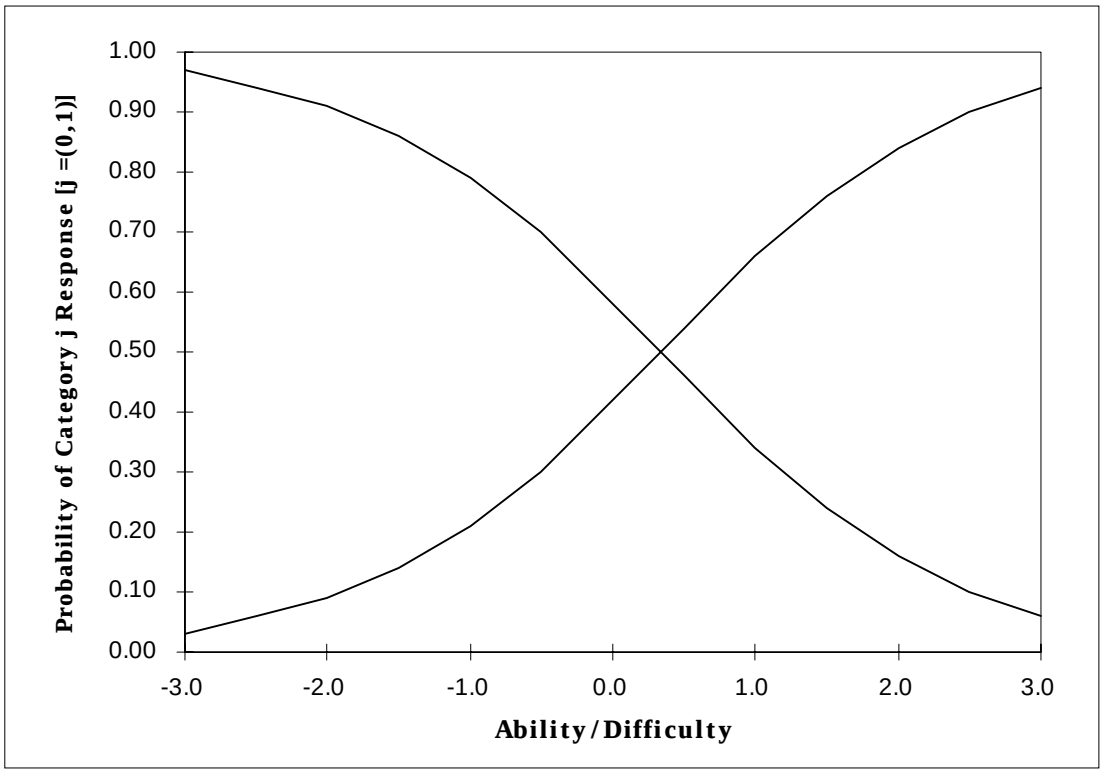

Figure 2 Category curves for a one-step item

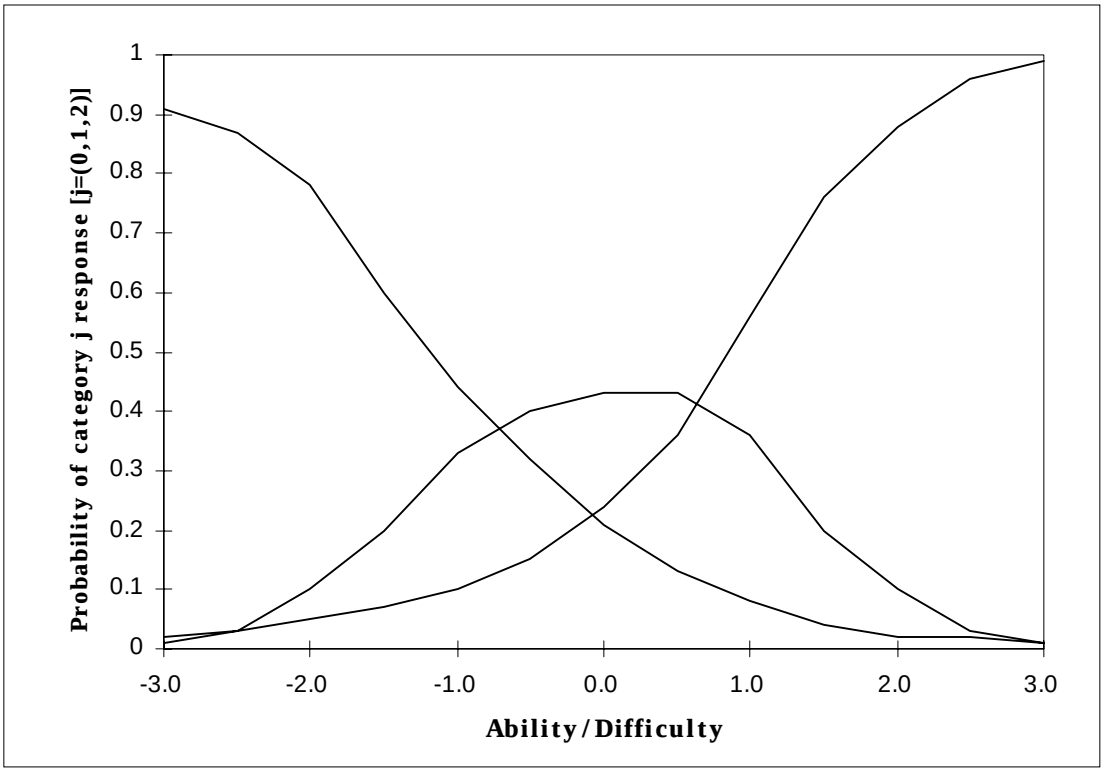

Figure 3 Category curves for a two-step item

Note that the likelihood of a category 2 response declines in both directions as ability decreases to the low extreme or increases to the high extreme. These points then may be thought of as the difficulties of crossing the "steps" between categories.

The most salient implication of the formulation can be summarized as follows. If the commonly used Rasch model applied to dichotomously (right/wrong) scored items can be thought of as simply a special case of the Masters Partial Credit model (applying to "one-step" items), then the act of scaling multiple-choice or "one-step" items together with "multi-step" items, whether they have two, three, or ten steps, is a straightforward process of applying the measurement model. The quality of the scaling then can be assessed in terms of known procedures.

For open-ended items that were not scored dichotomously (such as the *SOL* writing assessments), Harcourt Educational Measurement used the Masters Partial Credit Model. If the commonly used Rasch model applied to dichotomously (right/wrong) scored items can be thought of as simply a special case of the PCM (applying to "one-step" items), then the act of scaling multiple-choice or "one-step" items together with "multi-step" items, whether they have two, three, or ten steps, is a straightforward process of applying the measurement model.

One important property of the PCM is the separability of estimation of item/task parameters and person parameters. With the PCM, as with the Rasch model, the total score given by the sum of the categories in which a person responds is a sufficient statistic for estimating person ability (i.e., no additional information need be estimated). The total number of responses across examinees in a particular category is a sufficient statistic for estimating the step difficulty for that category. This is an important distinguishing feature of the PCM from other polytomous IRT

models, such as the Graded Response model (GRM) (Samejima, 1969) or other extensions of GRM in which person ability is estimated over all possible response patterns and item/task difficulties are weighted by item discrimination.

With PCM, the same total score will yield the same ability estimate for different examinees. With GRM, the same total raw score may yield different ability estimates, depending on the response patterns of the examinees ("pattern scoring"). In practical testing situations that involve the interpretation of scores on a test by the students, parents, and teachers, it is desirable for the same total score to have the same meaning. The PCM is the only measurement model allowing for such interpretation.

Sensitivity is another useful characteristic of the PCM. The Rasch model and its extensions are more sensitive to departure from unidimensionality than other polytomous models. For Rasch model proponents, significant variation of item discrimination is indicative of a dimensionality problem, rather than a purely psychometric phenomenon. Significant variation in item/task discrimination implies that the items are not rank-ordering examinees in the same way they should for a unidimensional instrument. The Rasch model and the PCM identify as misfitting an item with a significant departure from the expected level of discrimination so that judgments can be made regarding the extent to which that element of the assessment fairly measures student performance.

The PCM is a direct extension of the dichotomous one-parameter item response theory (IRT) model developed by Rasch in the 1950s (Rasch, 1980). For an item/task involving *m* score categories, one general expression for the probability of person *n* scoring *x* on item/task *i* is given by

$$P_{nxi} = \exp\sum_{j=0}^{x}\left(B_n - D_{ij}\right) / \sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}\left(B_n - D_{ij}\right)$$

where $x = 0,1,\ldots m$, and by definition,

$$\sum_{j=0}^{0}\left(B_n - D_{ij}\right) = 0$$

The above equation gives "the probability of scoring *x* on the *m*-th step of item/task *i* as a function of the person's position B_n on the variable (i.e., ability) and the difficulty of the *m* steps of item/task *i*" (Masters, 1982).

According to this model, the probability of an examinee scoring in a particular category (step) is the sum of the logit (log-odds) differences between Bn and Dij of all the completed steps, divided by the sum of the differences of all the steps of an item. Thissen and Steinberg (1986) refer to this model as a divide-by-total model. The parameters estimated by this model are (1) an ability estimate for each person (or ability estimate at each raw score level) and (2) *m* step (difficulty) estimates for each item/task with $m + 1$ score categories.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education.

Camilli, G. & Shepard, L.A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: SAGE Publications.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt Rinehart Wilson.

Haertel, E. H. (1996). *Estimating the decision consistency from a single administration of a performance assessment battery. A report on the National Board of Professional Teaching Standards McGEN Assessment*. Palo Alto, CA: Stanford University.

Linacre, J. M. & Wright, B. D. (1991). *BIGSTEPS*. (Rasch model computer program). Chicago, IL: MESA Press.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-1987.

Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8,* 452-461.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica, 47,* 149-174.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Rentz, R. R. (1980). *TRIAN*. (Rasch model and traditional statistics computer program). Athens, GA: Georgia State University.

Reckase, M. D. (2000, June). *The evolution of the NAEP Achievement Levels Setting Process: A summary of the research and development efforts conducted by ACT.* (A report for the National Assessment Governing Board, Washington, DC). Iowa City, IA: ACT, Inc.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 17*.

SAS Institute. (1989). *SAS System: Version 6.08*. (General purpose statistical software system). Cary, NC: The SAS Institute, Inc.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrica, 51,* 567-577.

Virginia Department of Education. (1999, February). *Standards of Learning (SOL) tests validity and reliability information: Spring 1998 administration*. (Report). Virginia Department of Education. Division of Assessment and Reporting. Richmond, VA: Author.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Young, M. J. & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment.* (CSE Technical Report 475). Center

for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, CA: University of California, Los Angeles.